

---

# Robustness in Both Domains: CLIP Needs a Robust Text Encoder

---

Elias Abad Rocamora<sup>EPFL</sup>, Christian Schlarman<sup>uni</sup>, Naman Deep Singh<sup>uni</sup>,  
Yongtao Wu<sup>EPFL</sup>, Matthias Hein<sup>uni</sup>, Volkan Cevher<sup>EPFL</sup>

<sup>EPFL</sup>: LIONS - École Polytechnique Fédérale de Lausanne, Switzerland

<sup>uni</sup>: Tübingen AI center, University of Tübingen, Germany  
{name.surname}@{epfl.ch, uni-tuebingen.de}

## Abstract

Adversarial input attacks can cause a significant shift of CLIP embeddings. This can affect the downstream robustness of models incorporating CLIP in the pipeline, such as text-to-image generative models or large vision language models. While some efforts have been done towards making the CLIP image encoders robust, the robustness of text encoders remains unexplored. In this work, we cover this gap in the literature. We propose LEAF: an efficient adversarial finetuning method for the text domain, with the ability to scale to large CLIP models. Our models significantly improve the zero-shot adversarial accuracy in the text domain, while maintaining the vision performance provided by robust image encoders. When combined with text-to-image diffusion models, we can improve the generation quality under adversarial noise. In multimodal retrieval tasks, LEAF improves the recall under adversarial noise over standard CLIP models. Finally, we show that robust text encoders facilitate better reconstruction of input text from its embedding via direct optimization. We open-source our code and models.

## 1 Introduction

Contrastive Language-Image Pretraining (CLIP) models embed images and captions into a shared embedding space [Radford et al., 2021]. CLIP is a simple but rather powerful tool for vision-language understanding, being employed in a wide range of multimodal tasks such as retrieval [Fang et al., 2021, Koukounas et al., 2024, Vendrow et al., 2024], Large Multimodal Models (LMMs) [Alayrac et al., 2022, Liu et al., 2023] and text-to-image generative models [Ramesh et al., 2021, Rombach et al., 2022, Ramesh et al., 2022, Podell et al., 2024].

However, the simplicity of CLIP and its plug-and-play usage becomes a double-edged sword, allowing adversarial attacks to be optimized over CLIP, and transferred to the downstream task of interest [Zhuang et al., 2023, Ghazanfari et al., 2023, 2024, Croce et al., 2025]. Recently, making the image encoder of CLIP robust has gained interest [Mao et al., 2023], making LMMs robust to adversarial perturbations by replacing the image encoder with an adversarially finetuned one [Schlarman et al., 2024]. Nevertheless, adversarial finetuning has not been yet investigated for the text encoder.

In this work, we fill this gap by studying adversarial finetuning for CLIP text encoders, proposing *Levenshtein Efficient Adversarial Finetuning* (LEAF). Motivated by recent advancements in the image domain, we optimize the same objective as Schlarman et al. [2024], allowing us to replace the text encoder in tasks like text-to-image generation, without needing to finetune the rest of the pipeline. Moreover, to make adversarial finetuning faster in the text domain, we propose an attack that can be parallelized within training batches, accelerating the approach of Abad Rocamora et al. [2024] by an order of magnitude with very little loss of performance.

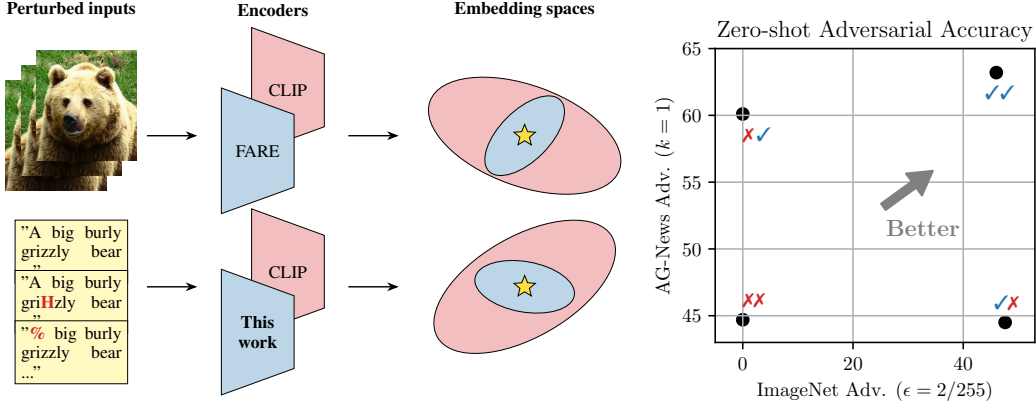


Figure 1: **Left: our idea.** Schlarman et al. [2024] propose FARE: finetuning the CLIP image encoder to produce embeddings close to the clean image embedding (★) under image perturbations. Analogously, we finetune the CLIP *text* encoder to produce embeddings close to the clean *text* embedding (★) under *text* perturbations. **Right: results in ViT-L/14.** The first (second)  $\times/\checkmark$  denotes the usage of a robust image (text) encoder. We constrain the text attacks with the Levenshtein distance and the image attacks in the  $\ell_\infty$  norm. By combining the FARE robust image encoder with our robust text encoder, we obtain high adversarial accuracy in both domains.

Our models, LEAF, are able to improve the zero-shot adversarial accuracy of CLIP models from 44.5% to 63.3% in AG-News at distance  $k = 1$  (one character change). When plugged into Stable Diffusion [Rombach et al., 2022, Podell et al., 2024], we achieve higher quality images under character-level perturbations. For retrieval tasks, our models achieve a recall 10 points higher on average than non-robust CLIP models at  $k = 2$ . Moreover, when inverting the embeddings of text encoders through direct optimization, we show that with LEAF models, we can recover a higher percentage of the original sentence. This results in LEAF encoders being more interpretable.

Overall, we show the robustness of CLIP text encoders can be improved with minimal effects on the clean performance in several tasks. We believe our robust CLIP models can make future models incorporating CLIP more robust and interpretable. Our code and models can be found in [github.com/LIONS-EPFL/LEAF](https://github.com/LIONS-EPFL/LEAF) and [huggingface.co/LEAF-CLIP](https://huggingface.co/LEAF-CLIP) respectively.

**Notation:** We use uppercase bold letters for matrices  $\mathbf{X} \in \mathbb{R}^{m \times n}$ , lowercase bold letters for vectors  $\mathbf{x} \in \mathbb{R}^m$  and lowercase letters for numbers  $x \in \mathbb{R}$ . Accordingly, the  $i^{\text{th}}$  row and the element in the  $i, j$  position of a matrix  $\mathbf{X}$  are given by  $\mathbf{x}_i$  and  $x_{ij}$  respectively. We use the operator  $|\cdot|$  for the size of sets, e.g.,  $|\mathcal{S}(\Gamma)|$  and the length of sequences, e.g., for  $\mathbf{X} \in \mathbb{R}^{m \times n}$ , we have  $|\mathbf{X}| = m$ . For two vectors  $\mathbf{u}, \mathbf{v} \in \mathbb{R}^h$ , we denote the cosine similarity as  $\text{sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{u}\|_2 \cdot \|\mathbf{v}\|_2}$ . We use the shorthand  $[n] = \{0, 1, \dots, n-1\}$  for any natural number  $n$ .

## 2 Background

In Section 2.1 we cover the approaches improving the adversarial robustness of CLIP. In Section 2.2 we discuss robustness in the text domain.

### 2.1 Robustness of CLIP

Let  $\mathcal{S}(\Gamma) = \{c_1 c_2 \dots c_m : c_i \in \Gamma \forall m \in \mathbb{N} \setminus \{0\}\}$  be the space of sequences of characters in the alphabet set  $\Gamma$ . We represent sentences  $\mathbf{S} \in \mathcal{S}(\Gamma)$  as sequences of one-hot vectors, i.e.,  $\mathbf{S} \in \{0, 1\}^{m \times |\Gamma|} : \|\mathbf{s}_i\|_1 = 1, \forall i \in [m]$ . Similarly, we can represent images with  $d$  pixels as real vectors  $\mathbf{x} \in \mathbb{R}^d$ . Overall, the training dataset is composed of  $n$  text-image pairs  $\{\mathbf{S}_i, \mathbf{x}_i\}_{i=1}^n$ .

The objective of CLIP is to learn a text encoder  $f_\theta : \mathcal{S}(\Gamma) \rightarrow \mathbb{R}^h$  and an image encoder  $g_\omega : \mathbb{R}^d \rightarrow \mathbb{R}^h$ , where  $h$  is the embedding size and  $\theta$  and  $\omega$  are the parameters of the text and image encoders respectively. Radford et al. [2021] propose to maximize the cosine similarity of positive

sentence-image pairs relative to the cosine similarity with other sentences and images in the dataset. We denote the weights obtained after pretraining with CLIP as  $\theta_{\text{CLIP}}$  and  $\omega_{\text{CLIP}}$ .

In order to make the image encoder  $g_\omega$  robust in the zero-shot classification task, Mao et al. [2023] use the sentences  $S_j = \text{“a photo of a LABEL}_j\text{,”}$   $\forall j \in [o]$ , where  $o$  is the number of classes. Then, given a dataset of images and labels  $\{x_i, y_i\}_{i=1}^n$ , so that  $y_i \in [o]$ , Mao et al. [2023] optimize:

$$\min_{\omega} \sum_{i=1}^n \max_{\|\delta_i\|_\infty \leq \epsilon} -\log \left( \frac{e^{\mathbf{f}_{\theta_{\text{CLIP}}}(S_{y_i})^\top g_\omega(x_i + \delta_i)}}}{\sum_{j=1}^o e^{\mathbf{f}_{\theta_{\text{CLIP}}}(S_j)^\top g_\omega(x_i + \delta_i)}}} \right). \quad (\text{TeCoA})$$

TeCoA significantly improves the robustness of the image encoder. However, it generalizes poorly to image classification tasks that are not part of the fine-tuning dataset, and degrades the performance when employed in an LMM pipeline, as shown by Schlarmann et al. [2024]. In order to overcome this, Schlarmann et al. [2024] propose FARE, which intends to preserve the original image embeddings while being robust. To do so, they optimize:

$$\min_{\omega} \sum_{i=1}^n \max_{\|\delta_i\|_\infty \leq \epsilon} \|g_{\omega_{\text{CLIP}}}(x_i) - g_\omega(x_i + \delta_i)\|_2^2. \quad (\text{FARE})$$

The FARE objective allows to employ the obtained image encoder within an LMM pipeline with minimal clean performance degradation. Motivated by these findings, in this work we construct a similar loss in the text domain (Eq. (TextFARE)) and adapt the algorithm to the challenges of this new domain (LEAF). See Fig. 1 for a visualization of the FARE and LEAF approaches.

## 2.2 Robustness in the text domain

Belinkov and Bisk [2018], Alzantot et al. [2018] showed that text classifiers are not robust to natural or adversarial noise, with text adversarial attacks being used in Large Language Models [Zou et al., 2023] and text-to-image generative models [Zhang et al., 2025]. Generally, given a sentence  $S$ , a model  $f$  and some loss function  $\mathcal{L}$ , the adversarial attack problem can be formulated as:

$$\max_{S' \in \mathcal{N}(S)} \mathcal{L}(f(S)),$$

where  $\mathcal{N}(S)$  is a set of neighboring sentences, i.e., the threat model. A great challenge in the text domain is defining a valid threat model, as the semantics of the sentence  $S$  should be preserved according to the task [Morris et al., 2020]. In the literature, we can categorize adversarial attacks into two main threat models: *token* and *character* level attacks. With token level attacks set to replace/insert/delete a small number of tokens in the sentence [Ren et al., 2019, Jin et al., 2020, Li et al., 2019, Garg and Ramakrishnan, 2020, Lee et al., 2022, Ebrahimi et al., 2018, Li et al., 2020, Guo et al., 2021, Hou et al., 2023]. Similarly, character-level attacks replace/insert/delete a small number of characters in the sentence [Belinkov and Bisk, 2018, Ebrahimi et al., 2018, Gao et al., 2018, Pruthi et al., 2019, Yang et al., 2020, Liu et al., 2022, Abad Rocamora et al., 2024]. Both approaches can be thought of as keeping a small Levenshtein distance [Levenshtein, 1966] between the original and attacked sentences in the token or character-level.

**Semantic constraints:** To ensure that semantics are preserved, token-level attacks usually constrain  $\mathcal{N}(S)$  further by only allowing token replacements between tokens with high similarity in the embedding space [Jin et al., 2020]. But, even with such semantic constraints, several works have pointed out that token level attacks do not preserve semantics [Morris et al., 2020, Dyrnishi et al., 2023], with Hou et al. [2023] reporting 56.5% of their attacks change the semantics of the sentence. Due to the difficulty in preserving semantics, we focus on character-level attacks in this work.

In the case of the character-level attacks, to further preserve semantics and simulate natural typos, some works constrain the attack to only replace characters that are nearby in the English keyboard [Belinkov and Bisk, 2018, Huang et al., 2019]. Others do not allow the attack to modify the first and last letter of words, to perturb short words, to perturb the same word twice or to insert special characters [Pruthi et al., 2019, Jones et al., 2020]. In the context of text-to-image generation, Chanakya et al. [2024] find that changing one character in the sentence can change one word for another and the text-to-image model accordingly generates a different object in the image. To avoid this, Chanakya et al. [2024] introduce the semantic constraint of not allowing new English words to appear after the attack. In this work, we decide to adopt the semantic constraints of [Chanakya et al., 2024] and find they are especially useful when performing adversarial finetuning of the CLIP text encoders, see Section 4.2.2.

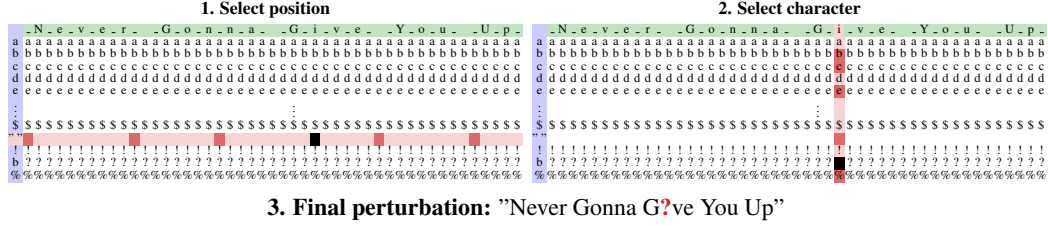


Figure 2: **Schematic and example of the attack used in LEAF:** In the first step, we randomly select  $\rho = 6$  positions, replace these with a whitespace and select the position with the highest loss. Next, we randomly select  $\rho$  characters from  $\Gamma$ , replace them in the chosen position and choose the one with the highest loss as the final perturbation. During training, the attack evaluates  $\rho \times B$  sentences in every forward pass, where  $B$  is the batch size. For more details, see Algorithm 1 in the appendix.

### 3 Method

In order to make the text encoder adversarially robust, we extend Eq. (FARE) to the text domain as:

$$\min_{\theta} \sum_{i=1}^n \max_{S'_i: d_{\text{Lev}}(S_i, S'_i) \leq k \wedge S'_i \in \mathcal{C}(S_i)} \|f_{\theta_{\text{CLIP}}}(S_i) - f_{\theta}(S'_i)\|_2^2, \quad (\text{TextFARE})$$

where the Levenshtein  $d_{\text{Lev}}$  distance is bounded by a parameter  $k$ , and  $\mathcal{C}(S)$  is either the complete set of sentences  $\mathcal{S}(\Gamma)$  or a subset only containing sentences with semantic constraints, see Section 2.2.

Intuitively, if the original CLIP encoder evaluated at the original sentence ( $f_{\theta_{\text{CLIP}}}(S)$ ) provides a good performance in downstream tasks, e.g., zero-shot classification or text-to-image generation, then, by solving Eq. (TextFARE), we will obtain a model that achieves similar performance under perturbations of the sentence. Moreover, Eqs. (FARE) and (TextFARE) allow for decoupled training of the text and image encoders.

Motivated by Danskin’s Theorem [Danskin, 1966, Latorre et al., 2023], we can (approximately) solve min-max problems by maximizing the inner problem and minimizing the error on the obtained perturbation. In the case of Eq. (FARE), Projected Gradient Descent (PGD) is used for the inner maximization problem [Madry et al., 2018, Schlarmann et al., 2024]. Similarly, we can use any adversarial attack to maximize the inner problem in Eq. (TextFARE), e.g., Gao et al. [2018], Abad Rocamora et al. [2024].

However, not every attack is adequate for adversarial finetuning, e.g., in the image domain, the strongest attacks in the AutoAttack ensemble [Croce and Hein, 2020] are never used during training due to their expensive time requirements. Contrarily, cheaper PGD attacks are used during training, providing fast training and generalization to stronger adversarial attacks Goodfellow et al. [2015], Madry et al. [2018], Shafahi et al. [2019], Wong et al. [2020]. The desiderata for an adversarial attack used during training can be captured by two points: (i) *High adversarial robustness to strong attacks after training*, (ii) *Low computational resources*.

As a baseline attack in the text domain, we select Charmer [Abad Rocamora et al., 2024]. Adversarial training with Charmer in text classification results in strong adversarial robustness, satisfying (i). Nevertheless, Charmer is not resource-efficient during training and thereby does not satisfy our second desiderata (ii). This is due to Charmer needing to evaluate a number of perturbations  $\mathcal{O}((2 \cdot |\mathcal{S}_i| + 1) + n_{\text{Charmer}} \cdot |\Gamma|)$ , which depends on the length of the sentence being attacked. This makes it harder to perform the attack simultaneously over sentences in a batch.

Overcoming this limitation, we propose *Levenshtein Efficient Adversarial Finetuning* (LEAF): utilizing a training-time attack that evaluates a constant number of perturbations  $\rho$  per sentence. Our attack replaces a test character (the whitespace) in  $\rho$  random positions within the sentence to select the position with the highest loss. Then,  $\rho$  random characters are replaced in the chosen position to choose again the one with the highest loss. Overall, this allows to perform the attack in two sequential evaluations of  $B \cdot \rho$  sentences, where  $B$  is the batch size. A visual representation of our attack is available in Fig. 2. Interestingly, if  $\rho = 1$ , our attack performs a random perturbation. For a more detailed discussion on LEAF, we refer to Appendix B. In Section 4.2 we empirically show LEAF satisfies our two desiderata.

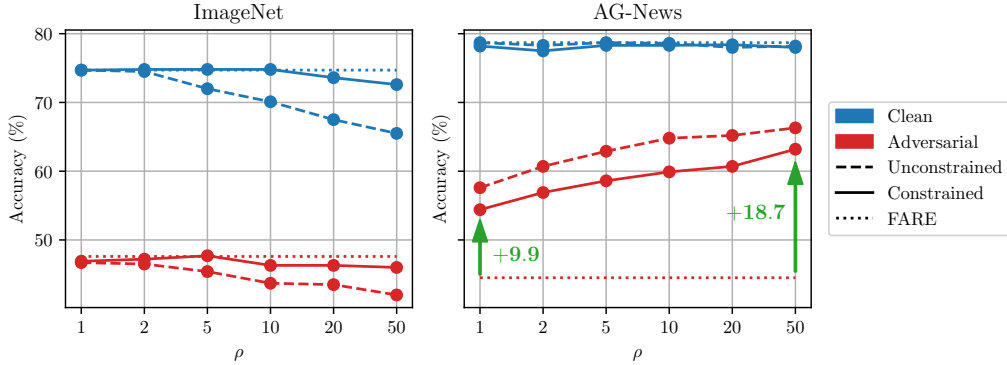


Figure 3: **Training hyperparameter effects:** We report the zero-shot clean and adversarial accuracy in the image (ImageNet) and text (AG-News) domains with FARE as a baseline. When no semantic constraints are employed (Section 2.2), the robustness in the text domain is improved at the cost of significantly degrading the image domain performance. Adding semantic constraints improves the robustness in the text domain with minimal effects on the image domain. Using random perturbations ( $\rho = 1$ ) improves the AG-News adversarial accuracy by 9.9 points, with stronger attacks ( $\rho = 50$ ) providing the best performance with 18.7 points of improvement.

## 4 Experiments

We start by introducing our experimental setup in Section 4.1. In Section 4.2 we cover our training results and display the interplay between  $\rho$ ,  $k$  and the usage of additional constraints during training. In Section 4.3 we present the performance of our models in zero-shot classification. In Section 4.4, we evaluate our CLIP models in multimodal retrieval tasks. In Section 4.5 we evaluate the performance of our CLIP text encoders when incorporated into text-to-image generative models. Finally, in Section 4.6 we evaluate how amenable our models are to embedding inversion. Additional experiments, including an evaluation with token-level attacks, are available in Appendix D.

### 4.1 Experimental setup

We train our text encoders for 30 epochs on the first 80,000 samples of the DataComp-small dataset [Gadre et al., 2023] with a batch size of 128 sentences,  $k = 1$ ,  $\rho = 50$  and semantic constraints, see Section 4.2.2, employing CLIP-ViT-L/14, OpenCLIP-ViT-H/14, OpenCLIP-ViT-g/14 and OpenCLIP-ViT-bigG/14 models. On the visual side, we scale the training method of Schlarman et al. [2024] to ViT-H/14 and ViT-g/14, using an  $\ell_\infty$  threat model with radius  $\epsilon = 2/255$ . See Appendix B.3 for a detailed account of hyperparameters. For evaluating the adversarial robustness with respect to image perturbations, we follow Schlarman et al. [2024] and employ the first two APGD attacks from the AutoAttack ensemble [Croce and Hein, 2020] with  $\epsilon = 2/255$ . In the text domain, we choose Charmer-20 with  $k = 1$  [Abad Rocamora et al., 2024] for evaluation. We employ the semantic constraints considered by [Chanakya et al., 2024] in the text-to-image and retrieval tasks. For the zero shot classification tasks, we do not employ such constraints as done by Abad Rocamora et al. [2024]. For a discussion on the use of constraints, we refer to Appendix D.1. For zero shot sentence classification with CLIP models, we follow the setup of Qin et al. [2023], see Appendix B.4 for more details. For additional details, we refer to Appendix D.

### 4.2 Training robust text encoders

In Section 4.2.1 we analyze the performance and training speed of Charmer and LEAF. In Section 4.2.2 we analyze how the performance is affected by our hyperparameters, i.e.,  $k$ ,  $\rho$  and  $\mathcal{C}(\mathcal{S})$ .

#### 4.2.1 Faster adversarial finetuning

First, we evaluate the performance of LEAF in terms of time and adversarial accuracy against training with Charmer [Abad Rocamora et al., 2024] with  $n_{\text{Charmer}} \in \{1, 20\}$ . To do so, we train CLIP-ViT-

Table 1: **Selecting the best attack for Adversarial Finetuning on ViT-B-32:** We measure the AG-News clean (Acc.) and adversarial accuracy (Adv.) at  $k = 1$  with Charmer-20 and the time in seconds to attack a batch of 128 sentences. We perform Adversarial Finetuning (Eq. (TextFARE)) for 1 epoch with  $k = 1$  using the attacks Charmer-1, Charmer-20 and LEAF with  $\rho \in \{20, 50\}$ . Our approach minimally affects the adversarial accuracy while being an order of magnitude faster than the fastest Charmer variant.

Defense	AG-News		Time (s)
	Acc. (%)	Adv. (%)	
Charmer-20	76.70( $\pm 0.14$ )	60.17( $\pm 0.31$ )	118.19( $\pm 53.68$ )
Charmer-1	76.37( $\pm 0.21$ )	<b>60.20</b> ( $\pm 0.37$ )	15.17( $\pm 28.98$ )
LEAF ( $\rho = 50$ )	76.63( $\pm 0.21$ )	59.80( $\pm 0.37$ )	3.23( $\pm 0.17$ )
LEAF ( $\rho = 20$ )	<b>76.87</b> ( $\pm 0.25$ )	58.30( $\pm 0.29$ )	<b>1.83</b> ( $\pm 0.11$ )

B-32 for 1 epoch at  $k = 1$  and using  $\rho \in \{20, 50\}$  for LEAF over three random training seeds. We measure the clean and adversarial accuracies with Charmer-20 on AG-News [Gulli, 2005, Zhang et al., 2015] and the average time to attack a batch of 128 samples.

In Table 1 we can observe that LEAF attains comparable clean and adversarial accuracies in comparison to the Charmer variants, while being significantly faster, i.e., 1.83 and 3.23 seconds per batch for our method in comparison to 15.17 and 118.19 seconds for the Charmer variants.

#### 4.2.2 The effect of our hyperparameters

In order to test the influence of our training hyperparameters, we finetune CLIP-ViT-L/14 initialized from pretrained FARE weights [Schlarmann et al., 2024] with  $\rho \in \{1, 2, 5, 10, 20, 50\}$ ,  $k \in \{1, 2\}$  and  $\mathcal{C}(S)$  including and not including semantic constraints. To evaluate how our method improves the robustness in the text domain, and affects the robustness in the image domain, we measure the clean and adversarial accuracies on ImageNet and AG-News.

In Fig. 3 we report the performance for  $k = 1$ . When increasing  $\rho$ , the adversarial accuracy in the text domain increases consistently. However, when employing unconstrained training attacks, both the clean and adversarial performance in the image domain are significantly degraded, e.g. at  $\rho = 50$ , a clean accuracy of 65.5% vs. 74.7% for the FARE model. In contrast, when applying semantic constraints, the improvements in robustness in the text domain follow a similar trend and the performance in the image domain is less degraded. For  $k = 2$ , we can extract the same insights, see Fig. 8. Overall, we select  $\rho = 50, k = 1$  and the use of semantic constraints during training.

#### 4.3 Zero-shot classification

We show the ImageNet and AG-News performance of the models when using robust encoders in image and/or text domain in Table 2 and Fig. 1. We observe that our robust text encoders introduce only minimal drop in image performance, while significantly improving the robustness in the text domain. Moreover, we observe that the effectiveness of FARE for fine-tuning robust image encoders that was demonstrated for ViT-L/14 by Schlarmann et al. [2024], extends to the larger ViT-H/14 and ViT-g/14 models. The lower performance of ViT-g/14 on ImageNet could be attributed to the smaller training batch size, see Appendix B.3. Importantly, only models that use a robust encoder in both domains achieve robustness in both tasks.

In Fig. 4 we report the adversarial accuracy of the ViT-L/14 sized models in the AG-News dataset for  $k \in \{0, 1, 2, 3, 4, 5\}$ , with  $k = 0$  representing the clean accuracy. Our model, while being trained with  $k = 1$ , is able to extrapolate the robustness to larger  $k$ . We observe that the CLIP and FARE models obtain a nearly zero adversarial accuracy for  $k \geq 4$ , while our model, is able to obtain the highest performance for any  $k$ .

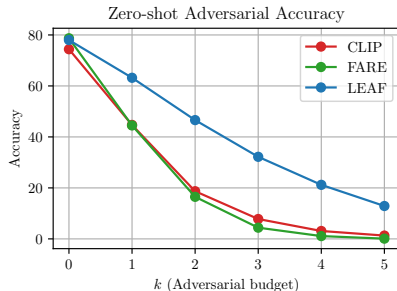


Figure 4: **Larger perturbations:** We evaluate the adversarial accuracy in AG-News for  $k \in \{1, 2, 3, 4, 5\}$  in the ViT-L/14 scale. Our model (LEAF) obtains the highest adversarial accuracy at all values of the distance bound  $k$ .

Table 2: **Zero-shot classification.** We report the adversarial accuracy (Adv.) on ImageNet with the first two attacks of AutoAttack (APGD-CE, APGD-t) at  $\epsilon = 2/255$  and on AG-News with Charmer-20 at  $k = 1$ . Only models employing robust image *and* text encoders are robust in both domains.

Robust Encoder		CLIP-ViT-L/14				OpenCLIP-ViT-H/14				OpenCLIP-ViT-g/14			
		ImageNet		AG-News		ImageNet		AG-News		ImageNet		AG-News	
Image	Text	Acc.	Adv.	Acc.	Adv.	Acc.	Adv.	Acc.	Adv.	Acc.	Adv.	Acc.	Adv.
✗	✗	76.4	0.0	74.4	44.7	77.2	0.0	71.1	37.6	77.8	0.0	67.3	35.8
✓	✗	74.7	47.6	78.7	44.5	76.8	48.4	70.7	37.5	73.8	41.8	66.4	32.9
✗	✓	73.4	0.0	73.9	60.1	77.0	0.0	71.1	50.2	76.3	0.0	67.3	47.4
✓	✓	72.6	46.0	78.0	63.2	76.8	46.3	72.3	53.3	72.0	41.3	66.7	46.3

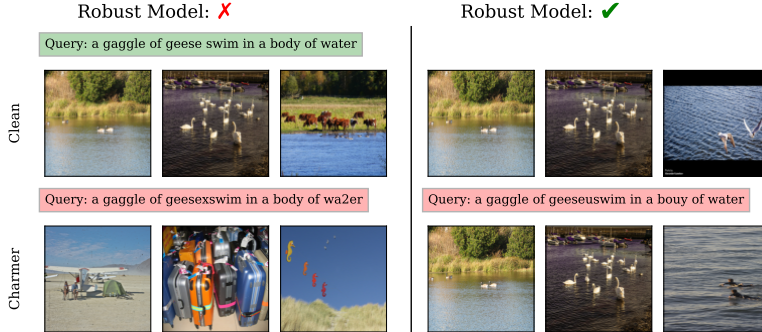


Figure 5: **Visualizing MS-COCO retrieved images.** For our ViT-L/14 robust model and its non-robust counterpart, we show the top-3 retrieved images for the original **Query** and the perturbed **Query** via Charmer ( $k = 2, n = 10$ ) attack. The robust model is able to preserve the order and retrieves semantically relevant images even for the perturbed query. More illustrations can be found in Appendix D.5. The target query in this case was “This is an image of a pyramid”.

#### 4.4 Text-image retrieval

Robustness of CLIP models to perturbations of textual queries is important as these models are often used as dataset/content filters Hong et al. [2024] and NSFW detectors Schuhmann et al. [2022], meaning any false negative can be detrimental. The robustness of retrieval based filters for visual adversaries has already been tested in Croce et al. [2025]. Consider the case where a CLIP based NSFW filter is queried with a perturbed query, any false negative retrieval here would be detrimental and concerning. To test how robust CLIP models are to such character based queries in retrieval setup, we test on the MS-COCO dataset as a proxy task.

For 1,000 validation set queries, the attack maximizes the similarity between the test query and a target string using different variants of the Charmer attack. Given some query text  $\mathcal{S}$  and corresponding embedding  $\mathbf{f}_\theta(\mathcal{S})$ , we maximize the cosine similarity between  $\mathbf{f}_\theta(\mathcal{S})$  and  $\mathbf{f}_\theta(\mathcal{T})$ , where  $\mathcal{T}$  is a target text semantically unrelated to  $\mathcal{S}$ . The objective takes the following form,

$$\max_{\mathcal{S}': d_{\text{lev}}(\mathcal{S}, \mathcal{S}') \leq k \wedge \mathcal{S}' \in \mathcal{C}(\mathcal{S})} \text{sim}(\mathbf{f}_\theta(\mathcal{S}'), \mathbf{f}_\theta(\mathcal{T})). \quad (1)$$

The optimization is done with the constrained Charmer attack for a different number of character changes.  $\mathcal{S}'$  is initialized with  $\mathcal{S}$ , and the overall perturbation set is constrained with  $\mathcal{C}(\mathcal{S})$  from Chanakya et al. [2024]. The formulation of the attack above can be seen as a targeted attack, the same attack can be done in an untargeted manner as in Eq. (2).

In Table 3, for different CLIP models, we show average *Recall* across 3 target strings, detailed results for each target can be found in Appendix D.5. For both 1 ( $k = 1$ ) and 2 ( $k = 2$ ) character perturbations, we see that the non-robust CLIP models retrieval performance goes down. Our robust models on the other hand showcase strong robustness while showing a small degradation in clean performance. For LEAF, the clean performance follows a trade-off with robustness depending on  $\rho$ , see Appendix D.5. Fig. 5, visualizes the attack and the top-3 retrieved images for a sample test query. Under perturbation, the non-robust model retrieves completely irrelevant images. The robust

Table 3: **MS-COCO text-to-image retrieval:** The statistics of the targeted Charmer adversarial attack (with  $k = 1, 2$  and semantic constraints) are averaged over 3 target strings.  $\times$ : denotes a non-robust CLIP model, whereas  $\checkmark$  indicates CLIP model robust in both image and text domains.

Model	Robust	Clean		Eval. $k$	Charmer-Con	
		Recall@1	Recall@5		Recall@1	Recall@5
CLIP-ViT-L/14	$\times$	49.11	73.79	1	37.31	62.67
	$\times$	49.11	73.79	2	30.66	52.76
	$\checkmark$	48.71	73.71	1	45.06	69.35
	$\checkmark$	48.71	73.71	2	40.22	65.09
OpenCLIP-ViT-H/14	$\times$	58.64	81.29	1	47.81	72.22
	$\times$	58.64	81.29	2	39.26	63.35
	$\checkmark$	56.80	80.65	1	52.97	77.26
	$\checkmark$	56.80	80.65	2	49.31	73.50
OpenCLIP-ViT-g/14	$\times$	60.64	82.22	1	47.93	72.71
	$\times$	60.64	82.22	2	37.51	61.82
	$\checkmark$	55.98	79.33	1	52.30	76.95
	$\checkmark$	55.98	79.33	2	48.71	73.71

model on the other hand, preserves the order and retrieves images relevant to the query. Moreover, in almost all cases it retrieves the top-1 image correctly, see Appendix D.5 for more such examples. Starting with  $k = 1$  text perturbations, we test the robustness of different variants of CLIP-ViT-L/14 models to bimodal attacks using APGD for image perturbations. Even in this more challenging setup, LEAF attains the most robust models, without sacrificing clean performance. We defer the associated results and discussion to Appendix D.5.1.

#### 4.5 Robustness of text-to-image models

In this section, we evaluate the performance of our robust text encoders when plugged into text-to-image generation pipelines. We take SD-1.5 [Rombach et al., 2022] and SDXL [Podell et al., 2024]. SD-1.5 employs the text encoder from ViT-L/14 and SDXL employs two text encoders: from ViT-L/14 and ViT-bigG/14. In order to attack the model, we follow Zhuang et al. [2023] by only accessing the text encoder. Given a sentence  $S$ , we employ Charmer-20 to solve:

$$\min_{S': d_{\text{Lev}}(S, S') \leq k \wedge S' \in \mathcal{C}(S)} \text{sim}(f_{\theta}(S), f_{\theta}(S')). \quad (2)$$

By minimizing the similarity between the original and perturbed embedding, we expect that the model generates images that do not align to the original caption. For SDXL, we maximize the average dissimilarities for both encoders. To analyze the quality of the generated images, through CLIP-ViT-B-16, we measure the CLIPScore between the original caption  $S$  and the generated image. In Fig. 6 we present the MS-COCO [Lin et al., 2014] SDXL image generation results. We can observe that the CLIPScore of SDXL with the LEAF encoders is significantly larger than the original SDXL for  $k \geq 1$ . On the right-hand-side of Fig. 6 we present the generated images for the first five captions in the MS-COCO validation dataset at  $k = 2$ , where for two captions, the original SDXL model produces completely different images compared to the original ones.

In Appendix D.3 we include additional text-to-image generation details and experiments over SD-1.5 and FLUX.1-dev [Black Forest Labs et al., 2025]. Interestingly, the generation quality of FLUX.1-dev can be severely degraded when only attacking its CLIP ViT-L/14 text encoder, see Table 13. We observe that the most common attack when the word "woman" appears, consists of replacing the final "n" for another character, see Table 19. This leads FLUX.1-dev to produce images of snakes as the tokens of the word "woma", a python species (Woma python), appear in the sentence. In Fig. 7 we report the images generated with FLUX.1-dev with the original CLIP encoder and the LEAF counterpart over 10 random seeds. When using our text encoder, the model is able to distinguish based on the rest of the sentence, whether a "woman" or a "woma" should be generated.



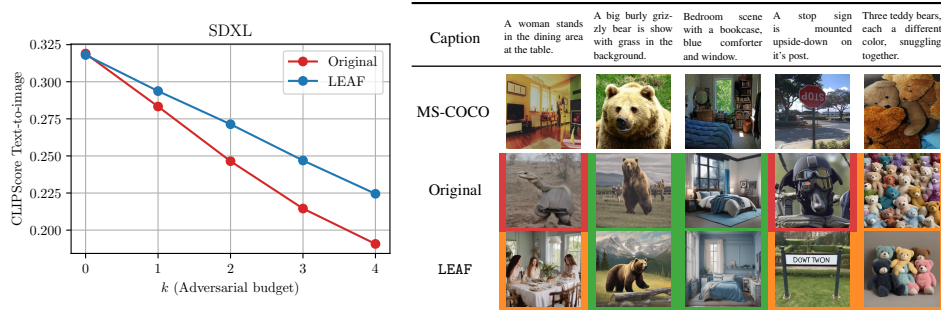


Figure 6: **Text-to-image generation results on SDXL:** On the left side, we present the MS-COCO CLIPScores of SDXL. The LEAF text encoders consistently improve the generation quality of SDXL under adversarial noise. On the right, we present the first five MS-COCO samples from the validation set and the corresponding SDXL generations at  $k = 2$ . The color borders indicate **null**, **partial** and **total** matching to the original image. With the original encoder, images 1 and 4 do not match at all the original ones. With the FARE encoders, all of the five images resemble the original ones, with some errors like the mismatch in the number of objects in image 5.

Table 4: **Text embedding inversion.** We invert text embeddings and measure the quality of reconstructions with various metrics. Robust models yield better reconstructions according to all metrics.

Model	Robust	sim $\uparrow$	Word Rec. $\uparrow$	Token Rec. $\uparrow$	BLEU $\uparrow$
CLIP-ViT-L/14	$\times$	0.89	34.4	38.9	8.3
	$\checkmark$	0.95	46.4	52.0	12.2
OpenCLIP-ViT-H/14	$\times$	0.86	33.5	34.1	8.9
	$\checkmark$	0.93	49.0	50.3	13.7
OpenCLIP-ViT-g/14	$\times$	0.94	43.7	48.1	5.6
	$\checkmark$	0.96	54.8	60.6	12.2

#### 4.6 Text embedding inversion

It is well known that robust models in the vision domain possess more interpretable gradients than clean models [Santurkar et al., 2019], which can be exploited to generate visual counterfactual explanations [Augustin et al., 2020, Boreiko et al., 2022]. Moreover, this allows to reconstruct images from their embeddings of a robust model by direct gradient based optimization [Croce et al., 2025].

We test if this advantageous property of robust vision models also holds in robust text models. To this end, we study the ability to invert text embeddings. Given an embedding  $f_{\theta}(S)$ , the goal is to reconstruct the unknown text  $S$ . Therefore we aim to solve the objective

$$\max_{S' \in \mathcal{S}(\Gamma)} \text{sim}(f_{\theta}(S'), f_{\theta}(S)). \quad (3)$$

To this end, we use the optimization method from Wen et al. [2023], where the text is initialized uniformly at random over the vocabulary of tokens and optimized via a gradient based algorithm.

We randomly sample 100 captions from MS-COCO, embed them via the given original and robust text encoders, and measure the success of reconstruction with four metrics: The cosine similarity between  $f_{\theta}(S')$  and  $f_{\theta}(S)$ , i.e., the objective in Eq. (3). *Word Recall* and *Token Recall* are the percentages of words/tokens in the original text that appear in the reconstruction, irrespective of order. Finally, BLEU [Papineni et al., 2002] is an ordering-aware similarity metric.

We show results in Table 4. The models with robust text encoders are best in every metric. Interestingly, we observe that the reconstructions of robust models generally improve when scaling up model size, while for non-robust models it does not improve from ViT-L/14 to ViT-H/14, but improves from ViT-H/14 to ViT-g/14. We observe that BLEU scores are low for all models, indicating that while many words are reconstructed correctly, their ordering is not. This could be attributed to the bag-of-words behavior of CLIP models discovered by Yüsekönül et al. [2023]. We show some randomly selected example reconstructions in Appendix Tables 22 and 23.





Caption	A woma@ stands in the xining area at the table.
Tokens	[ 'a</w>', 'wom', 'a</w>', '@</w>', 'stands</w>', 'in</w>', 'the</w>', 'x', 'ining</w>', 'area</w>', 'at</w>', 'the</w>', 'table</w>', '.</w>' ]
CLIP	
LEAF	
Caption	A woma python.
Tokens	[ 'a</w>', 'wom', 'a</w>', 'python</w>', '.</w>' ]
CLIP	
LEAF	

Figure 7: **Text-to-image generation with FLUX.1-dev:** We generate images with 10 random seeds using the original CLIP ViT-L/14 text encoder and the LEAF variant. The model using the CLIP text encoder consistently generates snakes for the first sentence, probably due to the appearance of the word "woma", a kind of snake (Woma python). When using our robust text encoder, we can accurately generate a woman and are also able to generate woma pythons when prompted to do so. While both captions start with ■, our text encoder distinguishes between the ■ and ■ continuations.

## 5 Conclusion

This work takes a first, systematic step toward *bimodal* robustness of CLIP by addressing the long-neglected text side. We introduced LEAF, a simple and efficient adversarial fine-tuning scheme for text encoders that mirrors the FARE philosophy on the image side: preserve the location of the clean embedding while enforcing invariance to small perturbations. For our adversarial fine-tuning scheme we develop a training-time character-level attack that allows for efficient training. In doing so, we showed that robustness in the text domain is both practically achievable and practically useful. Across zero-shot classification, text-to-image retrieval, and text-to-image generation, LEAF improves robustness to character-level attacks consistently, while leaving the clean performance intact.

Importantly, we show that robust CLIP text encoders obtained via LEAF can be combined with robust CLIP image encoders (e.g. FARE) to yield CLIP models that are robust on both input domains. This yields the first recipe that *jointly* elevates robustness in both modalities, and it scales without bespoke architectural changes or heavy joint training. Moreover, the method is modular: encoders can be swapped without touching downstream models, e.g. in text-to-image pipelines.

Notably, while we focus the empirical evaluation in this work on CLIP based models, our LEAF method could be applied to any text encoder: see Table 27 for an illustrative example beyond CLIP, where a BERT model is finetuned for sentence classification.

**Limitations:** Our robust image and text encoders are finetuned in isolation, joint training could yield larger robustness gains at higher training cost. Nevertheless, our bimodally robust models are validated against inference-time attacks that optimize over both modalities (see Table 25). In this work, we did not train models to be robust to token-level attacks, as these attacks often change the semantics of sentences [Dyrmishi et al., 2023]. Due to computational constraints, we did not train the largest image encoders (OpenCLIP-ViT-bigG) or the largest EVA-CLIP models [Sun et al., 2024]. Our approach has not yet been tested in other tasks using text encoders, e.g., RAG [Lewis et al., 2020]. We hope that our paper fosters advances in these areas.

## Acknowledgments

We thank the NeurIPS 2025 organization committee and reviewers for their work. This work was supported by the Swiss National Science Foundation (SNSF) under grant number 200021\_205011. Research was sponsored by the Army Research Office and was accomplished under Grant Number W911NF-24-1-0048. This work was supported by Hasler Foundation Program: Hasler Responsible AI (project number 21043). This work was supported as part of the Swiss AI Initiative by a grant from the Swiss National Supercomputing Centre (CSCS) under project ID a07 on Alps. EAR, YW and VC thank Gosia Baltain for her administrative help. We thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting CS and NDS. We acknowledge support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy (EXC number 2064/1, project number 390727645), as well as in the priority program SPP 2298, project number 464101476. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

## References

- Elias Abad Rocamora, Yongtao Wu, Fanghui Liu, Grigorios G. Chrysos, and Volkan Cevher. Revisiting character-level adversarial attacks for language models. In *International Conference on Machine Learning (ICML)*, 2024.
- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in neural information processing systems (NeurIPS)*, 2022.
- Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. Generating natural language adversarial examples. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.
- Maximilian Augustin, Alexander Meinke, and Matthias Hein. Adversarial robustness on in- and out-distribution improves explainability. In *ECCV*, 2020.
- Brian R Bartoldson, James Diffenderfer, Konstantinos Parasyris, and Bhavya Kailkhura. Adversarial robustness limits via scaling-law and human-alignment studies. In *International Conference on Machine Learning (ICML)*, pages 3046–3072, 2024.
- Yonatan Belinkov and Yonatan Bisk. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations (ICLR)*, 2018.
- Steven Bird and Edward Loper. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL <https://aclanthology.org/P04-3031/>.
- Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL <https://arxiv.org/abs/2506.15742>.
- Valentyn Boreiko, Maximilian Augustin, Francesco Croce, Philipp Berens, and Matthias Hein. Sparse visual counterfactual explanations in image space. In *GCPR*, 2022.
- Patibandla Chanakya, Putla Harsha, and Krishna Pratap Singh. Robustness of generative adversarial clips against single-character adversarial attacks in text-to-image generation. *IEEE Access*, 2024.

- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022. URL <https://arxiv.org/abs/2210.11416>.
- Mircea Cimpoi, Subhansu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014.
- Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*, 2011.
- Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning (ICML)*, 2020.
- Francesco Croce, Maksym Andriushchenko, Vikash Sehwal, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.
- Francesco Croce, Christian Schlarmann, Naman Deep Singh, and Matthias Hein. Adversarially robust clip models can induce better (robust) perceptual metrics. In *SaTML*, 2025.
- J. Danskin. The theory of max-min, with applications. *SIAM Journal on Applied Mathematics*, 14(4):641–664, 1966. doi: 10.1137/0114053.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019.
- Xinshuai Dong, Anh Tuan Luu, Rongrong Ji, and Hong Liu. Towards robustness against natural language word substitutions. In *International Conference on Learning Representations (ICLR)*, 2021.
- Salijona Dyrnishi, Salah Ghamizi, and Maxime Cordy. How do humans perceive adversarial text? a reality check on the validity and naturalness of word-based adversarial attacks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023.
- Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018.
- Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021.
- Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruva Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah M Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS)*, 2023.
- Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *IEEE Security and Privacy Workshops (SPW)*, 2018.

- Siddhant Garg and Goutham Ramakrishnan. BAE: BERT-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.
- Sara Ghazanfari, Siddharth Garg, Prashanth Krishnamurthy, Farshad Khorrami, and Alexandre Araujo. R-LPIPS: An adversarially robust perceptual similarity metric. In *ICML Workshop on New Frontiers in Adversarial Machine Learning*, 2023.
- Sara Ghazanfari, Alexandre Araujo, Prashanth Krishnamurthy, Farshad Khorrami, and Siddharth Garg. Lipsim: A provably robust perceptual similarity metric. In *ICLR*, 2024.
- Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *International Conference on Learning Representations (ICLR)*, 2015.
- Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy A Mann. Improving robustness using generated data. *Advances in neural information processing systems (NeurIPS)*, 34:4218–4233, 2021.
- Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007.
- Antonio Gulli. Ag’s corpus of news articles, 2005. URL [http://groups.di.unipi.it/~gulli/AG\\_corpus\\_of\\_news\\_articles.html](http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html).
- Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. Gradient-based adversarial attacks against text transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.
- Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7), 2019.
- Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 2021.
- Rachel Hong, William Agnew, Tadayoshi Kohno, and Jamie Morgenstern. Who’s in and who’s out? a case study of multimodal clip-filtering in datacomp. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 2024.
- Bairu Hou, Jinghan Jia, Yihua Zhang, Guanhua Zhang, Yang Zhang, Sijia Liu, and Shiyu Chang. Textgrad: Advancing robustness evaluation in NLP by gradient-driven optimization. In *International Conference on Learning Representations (ICLR)*, 2023.
- Po-Sen Huang, Robert Stanforth, Johannes Welbl, Chris Dyer, Dani Yogatama, Sven Gowal, Krishnamurthy Dvijotham, and Pushmeet Kohli. Achieving verified robustness to symbol substitutions via interval bound propagation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2019.
- Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL <https://doi.org/10.5281/zenodo.5143773>. If you use this software, please cite it as below.
- Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. *AAAI Conference on Artificial Intelligence*, 2020.
- Erik Jones, Robin Jia, Aditi Raghunathan, and Percy Liang. Robust encodings: A framework for combating adversarial typos. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.
- Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015.

- Andreas Koukounas, Georgios Mastrapas, Michael Günther, Bo Wang, Scott Martens, Isabelle Mohr, Saba Sturua, Mohammad Kalim Akram, Joan Fontanals Martínez, Saahil Ognawala, et al. Jina clip: Your clip model is also your text retriever. *arXiv preprint arXiv:2405.20204*, 2024.
- Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, 2013.
- Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Canada, 2009.
- Fabian Latorre, Igor Krawczuk, Leello Tadesse Dadi, Thomas Michaelson Pethick, and Volkan Cevher. Finding actual descent directions for adversarial training. In *International Conference on Learning Representations (ICLR)*, 2023.
- Deokjae Lee, Seungyong Moon, Junhyeok Lee, and Hyun Oh Song. Query-efficient and scalable black-box adversarial attacks on discrete sequential data via bayesian optimization. In *International Conference on Machine Learning (ICML)*, pages 12478–12497. PMLR, 2022.
- Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet Physics Doklady*, volume 10, pages 707–710. Soviet Union, 1966.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*, 2020.
- Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. Textbugger: Generating adversarial text against real-world applications. *Network and Distributed Systems Security (NDSS) Symposium*, 2019.
- Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.
- 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.
- Aiwei Liu, Honghai Yu, Xuming Hu, Shu’ang Li, Li Lin, Fukun Ma, Yawen Yang, and Lijie Wen. Character-level white-box adversarial attacks against transformers via attachable subwords substitution. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in neural information processing systems (NeurIPS)*, 2023.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.
- Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Association for Computational Linguistics (ACL)*, 2011.
- Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.
- Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft, 2013.
- Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. In *International Conference on Learning Representations (ICLR)*, 2023.

- Takeru Miyato, Andrew M. Dai, and Ian Goodfellow. Adversarial training methods for semi-supervised text classification. In *International Conference on Learning Representations (ICLR)*, 2017.
- John Morris, Eli Lifland, Jack Lanchantin, Yangfeng Ji, and Yanjun Qi. Reevaluating adversarial examples in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.
- John Xavier Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander M Rush. Text embeddings reveal (almost) as much as text. In *EMNLP*, 2023.
- John Xavier Morris, Wenting Zhao, Justin T Chiu, Vitaly Shmatikov, and Alexander M Rush. Language model inversion. In *ICLR*, 2024.
- Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*. IEEE, 2008.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.
- Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, 2012.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024.
- Samuele Poppi, Tobia Poppi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Safe-clip: Removing nsfw concepts from vision-and-language models. In *European Conference on Computer Vision (ECCV)*, pages 340–356. Springer, 2024.
- Danish Pruthi, Bhuwan Dhingra, and Zachary C. Lipton. Combating adversarial misspellings with robust word recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.
- Libo Qin, Weiyun Wang, Qiguang Chen, and Wanxiang Che. CLIPText: A new paradigm for zero-shot text classification. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL*, 2023.
- 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 (ICML)*. PMLR, 2021.
- Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning (ICML)*, 2021.
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.
- Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy Mann. Data augmentation can improve robustness. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in neural information processing systems (NeurIPS)*, 2021. URL <https://openreview.net/forum?id=kgVJBBThdSZ>.
- Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 10684–10695, 2022.

- Shibani Santurkar, Andrew Ilyas, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Image synthesis with a single (robust) classifier. In *NeurIPS*, 2019.
- Christian Schlarmann and Matthias Hein. On the adversarial robustness of multi-modal foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3677–3685, October 2023.
- Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust clip: Un-supervised adversarial fine-tuning of vision embeddings for robust large vision-language models. *International Conference on Machine Learning (ICML)*, 2024.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 2022.
- Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *Advances in neural information processing systems (NeurIPS)*, 2019.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013.
- Quan Sun, Jinsheng Wang, Qiyang Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. Eva-clip-18b: Scaling clip to 18 billion parameters, 2024. URL <https://arxiv.org/abs/2402.04252>.
- Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.
- Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations (ICLR)*, 2019. URL <https://openreview.net/forum?id=SyxAb30cY7>.
- Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *MICCAI*. Springer, 2018.
- Edward Vendrow, Omiros Pantazis, Alexander Shepard, Gabriel Brostow, Kate Jones, Oisín Mac Aodha, Sara Beery, and Grant Van Horn. Inquire: A natural world text-to-image retrieval benchmark. *Advances in neural information processing systems (NeurIPS)*, 37:126500–126514, 2024.
- Boxin Wang, Shuohang Wang, Yu Cheng, Zhe Gan, Ruoxi Jia, Bo Li, and Jingjing Liu. Info{bert}: Improving robustness of language models from an information theoretic perspective. In *International Conference on Learning Representations (ICLR)*, 2021.
- Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019.
- Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training. In *International Conference on Machine Learning (ICML)*, 2023.
- Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *NeurIPS*, 2023.
- Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *International Conference on Learning Representations (ICLR)*, 2020.
- Puyudi Yang, Jianbo Chen, Cho-Jui Hsieh, Jane-Ling Wang, and Michael I. Jordan. Greedy attack and gumbel attack: Generating adversarial examples for discrete data. *Journal of Machine Learning Research*, 21(43):1–36, 2020. URL <http://jmlr.org/papers/v21/19-569.html>.



- Yelp. Yelp open dataset, 2015. URL <https://business.yelp.com/data/resources/open-dataset/>.
- Mert Yükeşgönül, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *ICLR*, 2023.
- Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruysen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. A large-scale study of representation learning with the visual task adaptation benchmark, 2020. URL <https://arxiv.org/abs/1910.04867>.
- Chenyu Zhang, Mingwang Hu, Wenhui Li, and Lanjun Wang. Adversarial attacks and defenses on text-to-image diffusion models: A survey. *Information Fusion*, 114:102701, 2025.
- Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, 2019.
- 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.
- Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, 2015. URL [https://proceedings.neurips.cc/paper\\_files/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf).
- Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. Freelib: Enhanced adversarial training for natural language understanding. In *International Conference on Learning Representations (ICLR)*, 2020.
- Haomin Zhuang, Yihua Zhang, and Sijia Liu. A pilot study of query-free adversarial attack against stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2385–2392, 2023.
- 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.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [Yes]

Justification: We have a separate experimental section for each of the claims regarding the performance of our models exposed in the abstract.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We discuss our limitations at the end of Section 5.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate “Limitations” section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren’t acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not contain theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We cover our full set of hyperparameters in Section 4.1 and Appendix B.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

#### 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Our source code and documentation are published at <https://github.com/LIONS-EPFL/LEAF>.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We cover our hyperparameters in Section 4.1 and Appendix B.1. The analysis on the relevance of each hyperparameter is performed in Section 4.2.2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: While we report error bars for several experiments, it is not feasible for the large-scale training runs.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer “Yes” if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: As covered in Section 4.1, our text encoders were finetuned in a single NVIDIA A100 GPU with 40GB of memory. The largest image encoders required 8 40GB A100 GPUs. We did not exhaustively measure the time per experiment, but the largest finetuning run (ViT-bigG/14) took roughly 1 week to finish.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: Our research conducted in this paper conforms, in every respect, with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We cover the broader impact of our work in Appendix A

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We anticipate no such risk.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All of the models we finetune and the data we use is publicly available under MIT or cc-by-4.0 licenses. We appropriately credit the original authors by citing the corresponding papers and list the source models in Table 6.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

### 13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We give detailed descriptions of models trained in this work and will release them under MIT license.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

**16. Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.



## A Broader impact

This work positively impacts society by strengthening models that employ CLIP text encoders against perturbations in the text input, which is particularly important for safety-critical and high-volume applications. Practitioners can harden existing CLIP-based systems by adopting our adversarially robust text encoders as drop-in replacements with minimal changes. We provide source code and open source models to support responsible deployment.

## B Additional details

In this section, we provide additional details on the implementation of our method and the experimental setting.

**Additional Notation:** Given two matrices  $\mathbf{A} \in \mathbb{R}^{m \times d}$  and  $\mathbf{B} \in \mathbb{R}^{n \times d}$ , we define  $\mathbf{A} \oplus \mathbf{B} = \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix} \in \mathbb{R}^{(m+n) \times d}$ . Concatenating with the empty sequence  $\emptyset$  results in the identity  $\mathbf{A} \oplus \emptyset = \mathbf{A}$ . We denote as  $\mathbf{A}_2: \in \mathbb{R}^{(m-1) \times d}$  the matrix obtained by removing the first row.

### B.1 Method details

Firstly, we characterize the single-character perturbations following Abad Rocamora et al. [2024].

**Definition B.1** (Expansion and contraction operators). Let  $\mathcal{S}(\Gamma)$  be the space of sentences with alphabet  $\Gamma$  and the special character  $\xi \notin \Gamma$ , the pair of expansion-contraction functions  $\phi: \mathcal{S}(\Gamma) \rightarrow \mathcal{S}(\Gamma \cup \{\xi\})$  and  $\psi: \mathcal{S}(\Gamma \cup \{\xi\}) \rightarrow \mathcal{S}(\Gamma)$  is defined as:

$$\phi(\mathbf{S}) := \begin{cases} \xi & \text{if } |\mathbf{S}| = 0 \\ \xi, \mathbf{S}_1 \oplus \phi(\mathbf{S}_{2:}) & \text{otherwise.} \end{cases} \quad \psi(\mathbf{S}) := \begin{cases} \emptyset & \text{if } |\mathbf{S}| = 0 \\ \psi(\mathbf{S}_{2:}) & \text{if } \mathbf{S}_1 = \xi \\ \mathbf{S}_1 \oplus \psi(\mathbf{S}_{2:}) & \text{otherwise.} \end{cases}$$

Clearly,  $\phi(\mathbf{S})$  aims to insert  $\xi$  into  $\mathbf{S}$  in all possible positions between characters and at the beginning and end of the sentence, and thus we have  $|\phi(\mathbf{S})| = 2 \cdot |\mathbf{S}| + 1$ . Similarly,  $\psi(\mathbf{S})$  aims to remove all  $\xi$  occurred in  $\mathbf{S}$ . The  $(\phi, \psi)$  pair satisfies the property that  $\psi(\phi(\mathbf{S})) = \mathbf{S}$ . We give the following example for a better understanding.

*Example B.2.* Let  $\xi := \perp$  for visibility:

$$\phi(\text{Hello}) = \perp\text{H}\perp\text{e}\perp\perp\perp\perp\perp\text{o}\perp \quad \psi(\perp\text{H}\perp\text{e}\perp\perp\perp\perp\perp\text{o}\perp) = \text{Hello} \quad \psi(\perp\text{H}\perp\text{e}\perp\perp\perp\perp\perp\text{o}\perp) = \text{Helo} \quad \psi(\perp\text{H}\perp\text{e}\perp\perp\perp\perp) = \text{Hello}.$$

**Definition B.3** (Replacement operator). Let  $\mathbf{S} \in \mathcal{S}(\Gamma \cup \{\xi\})$ , the integer  $i \in [|\mathbf{S}|]$  and the character  $c$ , the replacement operator  $\overset{i}{\leftarrow} c$  of the  $i^{\text{th}}$  position of  $\mathbf{S}$  with  $c$  is defined as:

$$\mathbf{S} \overset{i}{\leftarrow} c := \mathbf{S}_{:i-1} \oplus c \oplus \mathbf{S}_{i+1:}$$

Thanks to Definition B.3, we are ready to present our attack in Algorithm 1. The advantage of Algorithm 1 resides in attacking a batch of  $B$  sentences in parallel, an important feature for efficient adversarial training.

### B.2 Semantic constraints details

In order to follow the semantic constraints of [Chanakya et al., 2024], we constrain the attacks during training and during retrieval and text-to-image generation to not produce new English words. To do so, we employ Algorithm 2 over pairs of sentences  $\mathbf{S}$  and  $\mathbf{S}'$  so that  $d_{\text{lev}}(\mathbf{S}, \mathbf{S}') = 1$ . Algorithm 2 returns that the perturbation  $\mathbf{S}'$  is valid only if it contains less english words than  $\mathbf{S}$ .

### B.3 Training details

All of our text encoders are trained on the first 80,000 samples of the DataComp-small dataset [Gadre et al., 2023] for 30 epochs with a batch size of 128 sentences. We employ the AdamW optimizer [Kingma and Ba, 2015, Loshchilov and Hutter, 2019], a weight decay of  $10^{-4}$ , a maximum learning rate of  $10^{-5}$  with a linear warmup of 1,400 steps and cosine decay. For training the robust

---

**Algorithm 1** LEAF batched attack

---

- 1: **Inputs:** Text encoder  $f_\theta : \mathcal{S}(\Gamma) \rightarrow \mathbb{R}^h$ , batch  $\{\mathcal{S}_i\}_{i=1}^B$ , loss function  $\mathcal{L}$ , radius  $k$ , number of simultaneous perturbations  $\rho$ , alphabet  $\Gamma$ , test character  $t$  and flag for semantic constraints Cons.
  - 2:  $\hat{\mathcal{S}}_i = \mathcal{S}_i \forall i \in [B]$  ▷ Initialize perturbations with clean sentences.
  - 3: **for**  $1, \dots, k$  **do**
  - 4:  $p_{ij} \sim \text{Unif.}([2 \cdot |\hat{\mathcal{S}}_i| + 1]) \forall i \in [B] \forall j \in [\rho]$  ▷ Sample  $\rho$  positions in every sentence.
  - 5:  $\bar{\mathcal{S}} = \left\{ \left\{ \psi \left( \phi(\hat{\mathcal{S}}_i) \stackrel{p_{ij}}{\leftarrow} t \right) \right\}_{j=1}^{\rho} \right\}_{i=1}^B$  ▷ Replace the test character in all  $p_{ij}$ .
  - 6: **if** Cons **then** ▷ Use Algorithm 2 to check if the perturbation is valid, revert otherwise.
  - 7:  $\bar{\mathcal{S}}_{ij} = \begin{cases} \bar{\mathcal{S}}_{ij} & \text{if valid}(\hat{\mathcal{S}}_i, \bar{\mathcal{S}}_{ij}) \\ \hat{\mathcal{S}}_i & \text{otherwise} \end{cases} \forall i \in [B] \forall j \in [\rho]$
  - 8:  $j_i^* = \arg \max_{j \in [\rho]} \mathcal{L}(f_\theta(\bar{\mathcal{S}}_{ij}))$  ▷ Eval. losses in parallel and get the max.
  - 9:  $c_{ij} \sim \text{Unif.}(\Gamma) \forall i \in [B] \forall j \in [\rho]$  ▷ Sample  $\rho$  characters for every sentence.
  - 10:  $\bar{\mathcal{S}} = \left\{ \left\{ \psi \left( \phi(\hat{\mathcal{S}}_i) \stackrel{p_{ij_i^*}}{\leftarrow} c_{ij} \right) \right\}_{j=1}^{\rho} \right\}_{i=1}^B$  ▷ Replace  $c_{ij}$  in the position  $p_{ij_i^*}$ .
  - 11: **if** Cons **then** ▷ Use Algorithm 2 to check if the perturbation is valid, revert otherwise.
  - 12:  $\bar{\mathcal{S}}_{ij} = \begin{cases} \bar{\mathcal{S}}_{ij} & \text{if valid}(\hat{\mathcal{S}}_i, \bar{\mathcal{S}}_{ij}) \\ \hat{\mathcal{S}}_i & \text{otherwise} \end{cases} \forall i \in [B] \forall j \in [\rho]$
  - 13:  $l_i^* = \arg \max_{j \in [\rho]} \mathcal{L}(f_\theta(\bar{\mathcal{S}}_{ij}))$  ▷ Eval. losses in parallel and get the max.
  - 14:  $\hat{\mathcal{S}}_i = \bar{\mathcal{S}}_{il_i^*} \forall i \in [B]$  ▷ Update perturbations.
  - 15: **return**  $\left\{ \hat{\mathcal{S}}_i \right\}_{i=1}^B$
- 

---

**Algorithm 2** Semantic constraints

---

- 1: **Inputs:** Sentence  $\mathcal{S}$  and perturbation  $\mathcal{S}'$ .
  - 2:  $m = |\text{words}(\mathcal{S})|$
  - 3:  $n = |\text{words}(\mathcal{S}')|$  ▷ We extract English words using NLTK: <https://www.nltk.org/>
  - 4: **return**  $m > n$
- 

vision encoder, we adapt the setup of Schlarmann et al. [2024]. Namely, we train on images from ImageNet for 10k steps (instead of 20k, due to compute constraints) with a batch size of 128 for ViT-H/14 and 64 for ViT-g/14. We use weight decay of  $10^{-4}$ , a maximum learning rate of  $10^{-5}$  with a linear warmup of 700 steps and cosine decay. To optimize the inner adversarial objective, we use PGD with 10 steps and set  $\epsilon = 2/255$ . Our codebase is based on OpenCLIP [Ilharco et al., 2021]. All of our experiments are conducted in a single Nvidia A100 40GB GPU, except for training robust image encoders, where 8 GPUs were employed.

#### B.4 Zero-shot text classification

Analogously to how zero-shot image classification is performed in the original CLIP paper [Radford et al., 2021], Qin et al. [2023] encode one image representing each class and compute the similarities with the sentence embedding. Then the predicted class is the one with the highest cosine similarity in the embedding space. In Table 5 we present the images employed for each dataset and label.

#### B.5 Text inversion

In order to invert text embeddings, we sample 100 random captions from COCO val2017 and use the optimization method proposed by Wen et al. [2023] with 3000 iterations, learning rate 0.1, and weight decay 0.1.

Table 5: Images and sentences used for zero-shot text classification.







Dataset	Images			
	Class 1	Class 2	Class 3	Class 4
SST-2 / IMDB / Yelp			NA	NA
AG-News				
	Sentences			
SST-2 / IMDB / Yelp	"Negative Review"	"Positive Review"	NA	NA
AG-News	"World News"	"Sports News"	"Business News"	"Science and Technology News"

Table 6: Source models employed for finetuning and evaluation.

Model	Source
CLIP-ViT-B-32	<a href="https://huggingface.co/openai/clip-vit-base-patch32">https://huggingface.co/openai/clip-vit-base-patch32</a>
CLIP-ViT-B-16	<a href="https://huggingface.co/openai/clip-vit-base-patch16">https://huggingface.co/openai/clip-vit-base-patch16</a>
ViT-L/14	<a href="https://huggingface.co/openai/clip-vit-large-patch14">https://huggingface.co/openai/clip-vit-large-patch14</a>
FARE	<a href="https://huggingface.co/chs20/fare2-clip">https://huggingface.co/chs20/fare2-clip</a>
SafeCLIP	<a href="https://huggingface.co/aimagelab/safeclip_vit-l_14">https://huggingface.co/aimagelab/safeclip_vit-l_14</a>
OpenCLIP-ViT-H-14	<a href="https://huggingface.co/laion/CLIP-ViT-H-14-laion2B-s32B-b79K">https://huggingface.co/laion/CLIP-ViT-H-14-laion2B-s32B-b79K</a>
OpenCLIP-ViT-g-14	<a href="https://huggingface.co/laion/CLIP-ViT-g-14-laion2B-s12B-b42K">https://huggingface.co/laion/CLIP-ViT-g-14-laion2B-s12B-b42K</a>
OpenCLIP-ViT-bigG-14	<a href="https://huggingface.co/laion/CLIP-ViT-bigG-14-laion2B-39B-b160k">https://huggingface.co/laion/CLIP-ViT-bigG-14-laion2B-39B-b160k</a>
Stable Diffusion v1.5 (SD-1.5)	<a href="https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5">https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5</a>
Stable Diffusion XL base v1.0 (SDXL)	<a href="https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0">https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0</a>
FLUX.1-dev	<a href="https://huggingface.co/black-forest-labs/FLUX.1-dev">https://huggingface.co/black-forest-labs/FLUX.1-dev</a>

## B.6 Model checkpoints

In Table 6, we enumerate the external models employed in this work and the sources used for comparison and finetuning.

## C Related work

In this section we cover related work on Adversarial Attacks, Adversarial Training, Robustness of Multimodal Models and text inversion.

**Adversarial Attacks** The vulnerability of deep learning models against adversarial input attacks is well known [Szegedy et al., 2014, Goodfellow et al., 2015] and has been extensively studied in the vision input domain [Croce and Hein, 2020, Schlarmann and Hein, 2023] and the text input domain, with the most popular attacks employing perturbations in the token-level [Ren et al., 2019, Jin et al., 2020, Li et al., 2019, Garg and Ramakrishnan, 2020, Lee et al., 2022, Ebrahimi et al., 2018, Li et al., 2020, Guo et al., 2021, Hou et al., 2023] and character-level [Belinkov and Bisk, 2018, Ebrahimi et al., 2018, Gao et al., 2018, Pruthi et al., 2019, Yang et al., 2020, Liu et al., 2022, Abad Rocamora et al., 2024].

**Adversarial Training in the text domain.** Adversarial Training [Madry et al., 2018] and its variants [Zhang et al., 2019, Rebuffi et al., 2021, Goyal et al., 2021, Wang et al., 2023, Bartoldson et al., 2024] are the most prominent defense against adversarial examples in the image domain Croce and Hein [2020], Croce et al. [2020].

In the text domain, also variants of adversarial training constitute the best defenses, with most defenses focusing on token-level attacks. Taking advantage of the efficiency of PGD, Miyato et al. [2017] propose solving the inner maximization problem in a  $\ell_p$  constrained ball around every token embedding. Zhu et al. [2020] accelerate embedding-level PGD AT and show improvements in clean accuracy. Wang et al. [2021] show improvements in adversarial accuracy by adding an information theoretic regularization term. Deviating from the embedding-based PGD AT paradigm, Dong et al. [2021] use PGD to maximize the loss over a convex combination of synonym embeddings. Then, Hou et al. [2023] find that directly optimizing the inner max in the text space with existing attacks [Jin et al., 2020] significantly boosts the adversarial accuracy against multiple adversarial attacks.

In the character-level, it was initially thought that typo-correctors would suffice as a defense [Pruthi et al., 2019, Jones et al., 2020]. Abad Rocamora et al. [2024] shows that typo-corrector defenses can be easily broken. Additionally Abad Rocamora et al. [2024] show that similarly to the results of [Hou et al., 2023] in the token-level, performing adversarial training with character-level perturbations improved the character-level robustness.

**Robustness of Multimodal Models.** Attacking and defending multimodal models has gained significant interest recently. Mao et al. [2023] propose TeCoA, which performs supervised adversarial fine-tuning on CLIP in order to defend against visual adversarial attacks. In turn, Schlarman et al. [2024] propose FARE, an unsupervised robust fine-tuning method for vision encoders that preserves downstream performance, e.g. of LMMs that utilize a vision encoder.

**Text inversions.** Morris et al. [2023, 2024] learn models that can invert text embeddings or language model outputs. In contrast, Wen et al. [2023] invert CLIP image embeddings into text via direct optimization. They use the reconstructed text to prompt diffusion models and thereby generate similar images. We use their optimization scheme to invert text embeddings and show that it yields better results when used with our robust models.

## D Additional experiments

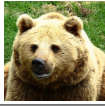







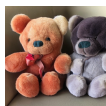





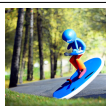
In this section we cover additional experiments not fitting in the main manuscript. First, in Appendix D.1, we analyze the effect adding additional constrains to the adversarial attack. Then, in Appendix D.2 we cover additional experiments in zero-shot classification. In Appendix D.3 we include additional text-to-image generation experiments. In Appendix D.4 we include examples of the sentences reconstructed from their embeddings through embedding inversion. Finally, In Appendix D.6, we perform ablations studying the final losses for different values of  $k$  and  $\epsilon$ , and perform token-level adversarial attacks.

### D.1 On the effect of additional attack constrains for Text-to-image models

In this section, we evaluate the effectiveness of the semantic constraints considered by Chanakya et al. [2024]. In order to avoid including new words with different information in the prompt, Chanakya et al. [2024] constrain the attack to not produce new words in the English vocabulary. To do so, they tokenize the clean and adversarial prompts and check for the appearance of new words in the adversarial prompt based on the NLTK English dictionary [Bird and Loper, 2004]. In order to check for the need of these constraints, we attack SD-1.5 equipped with our robust text encoder at  $k = 2$  using Charmer [Abad Rocamora et al., 2024] on the COCO val2017 dataset [Lin et al., 2014]. We then visually explore the adversarial prompts and generated images to look for inconsistencies.

In Table 7 we can observe five examples of unconstrained attacks producing adversarial prompts with significantly different meaning. Since the only constraint is that the Levenshtein distance needs to be  $\leq 2$ , the attack is able to turn "bear" into "beer", "stop" into "shop", "bananas" into "bandanas" or "wave" into "pave". This results in the diffusion model generating images that correctly adopt these adversarial captions and the adversarial prompts being invalid. If we constrain the attacker to not generate new words, the adversarial prompts preserve the meaning of the original captions up to uncommon words/abbreviations not present in the NLTK dictionary, like "grads" or "smurfs". Overall, we consider the constraints necessary for the text-to-image generation tasks, agreeing with Chanakya et al. [2024].

Table 7: **Examples of problematic attacks in COCO val2017:** If no additional constraints are considered, a single character change can produce semantical changes in the prompt, e.g., "bear" is transformed into "beer". This leads to image generations that are highly dissimilar to the original reference image, but are correct according to the adversarial prompt. The semantic constraints employed by Chanakya et al. [2024] help reducing the amount of new words. Nevertheless, some abbreviations like "grads" or uncommon words like "smurf" still appear after the attack.

ID	Original caption	Original image	Unconstrained		Constrained [Chanakya et al., 2024]	
			Adversarial caption	Generated image	Adversarial caption	Generated image
285	A big burly grizzly bear is show with grass in the background.		A big burly grizzly <b>beer</b> is show with <b>brass</b> in the background.		A big burly !rizzly bear is show with <b>grads</b> in the background.	
724	A stop sign is mounted upside-down on it's post.		A <b>shop</b> sign is mounted up!side-down on it's post.		A scop sign is mountedaupside-down on it's post.	
776	"Three teddy bears, each a different color, snuggling together."		"Tree teddy <b>beans</b> , each a different color, snuggling together."		8hree teddy bears, each a different color, snuggling toge,ther.	
3661	A bunch of bananas sitting on top of a wooden table.		A bunch of <b>bandanas</b> sitting on top of aawooden table.		A bunch of bananas sitti-g on top of a woodenitable.	
6460	a person riding a surf board on a wave		a person riding a <b>smurf</b> board on a <b>pave</b>		a person riding a <b>smurf</b> board on a waze	

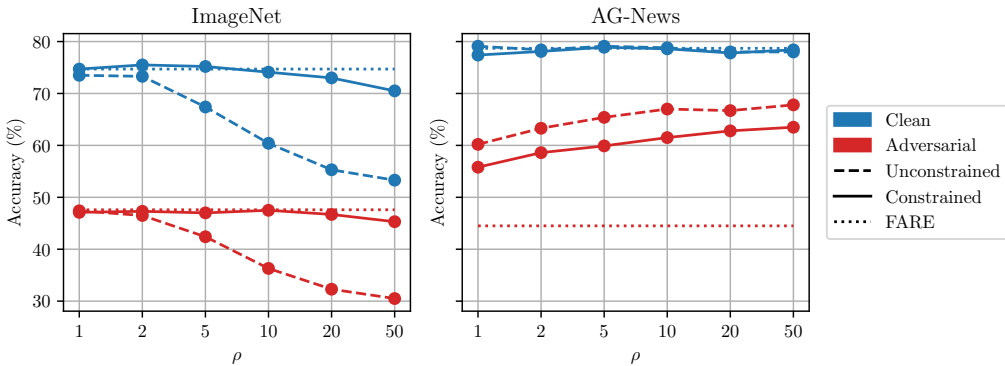


Figure 8: **Hyperparameter effects at  $k = 2$ :** We report the zero-shot clean and adversarial accuracy in both domains (ImageNet and AG-News) with FARE [Schlarmann et al., 2024] as a baseline. For the unconstrained attack, larger values of  $\rho$  improve the robustness in the text domain at the cost of significantly degrading the clean and adversarial performance in the image domain. Constraining the attack allows improving the robustness in the text domain with minimal effects on the image domain performance.

## D.2 Zero-shot classification

In this section we include additional datasets for zero-shot image and text classification. We also include a hyperparameter analysis with  $k = 2$ .

In Fig. 8 we can observe the same experiment as in Section 4.2.2 and Fig. 3 with  $k = 2$  instead of  $k = 1$ . Similarly to the experiments with  $k = 1$ , increasing  $\rho$  leads to a degraded performance in the image domain when no constraints are employed. Including the constraints, allows for increasing

Table 8: **Zero-shot performance for different  $k$ ,  $\rho$  and constraints.**

Semantic Constraints	$k$	$\rho$	ImageNet		AG-News	
			Acc.	PGD-20 Acc. ( $\epsilon = \frac{2}{255}$ )	Acc.	Charmer Acc. ( $k = 1$ )
✗	1	1	74.7	46.7	78.7	57.6
		2	74.5	46.5	78.3	60.7
		5	72.0	45.4	78.7	62.9
		10	70.1	43.7	78.6	64.8
		20	67.5	43.5	78.0	65.2
		50	65.5	42.0	78.2	66.3
	2	1	73.5	47.4	79.1	60.2
		2	73.3	46.5	78.4	63.3
		5	67.4	42.4	79.1	65.4
		10	60.4	36.3	78.8	67.0
		20	55.3	32.3	78.0	66.7
		50	53.3	30.5	78.0	67.8
✓	1	1	74.7	46.9	78.2	54.4
		2	74.8	47.2	77.5	56.9
		5	74.8	47.7	78.3	58.6
		10	74.8	46.3	78.3	59.9
		20	73.6	46.3	78.4	60.7
		50	72.6	46.0	78.0	63.2
	2	1	74.7	47.1	77.4	55.8
		2	75.5	47.3	78.1	58.6
		5	75.2	47.0	78.9	59.9
		10	74.1	47.5	78.6	61.5
		20	73.0	46.7	77.8	62.8
		50	70.5	45.3	78.4	63.5

the robustness in the text domain with less performance degradation. The numbers from Figs. 3 and 8 are available in Table 8.

### D.2.1 Additional experiments on zero-shot image classification

For zero-shot image classification, we measure the clean and robust accuracy on 13 datasets: CalTech101 Griffin et al. [2007], StanfordCars Krause et al. [2013], CIFAR10, CIFAR100 Krizhevsky [2009], DTD Cimpoi et al. [2014], EuroSAT Helber et al. [2019], FGVC Aircrafts Maji et al. [2013], Flowers Nilsback and Zisserman [2008], ImageNet-R Hendrycks et al. [2021], ImageNet-Sketch Wang et al. [2019], PCAM Veeling et al. [2018], OxfordPets Parkhi et al. [2012], and STL-10 Coates et al. [2011]. To measure robustness, we conduct visual attacks as described in Section 4.1, and restrict the evaluation to 1000 random samples on all datasets. We evaluate original models and models that employ robust encoders in both domains. Results are reported in Table 9. The robust models maintain much better performance under adversarial attacks, while sacrificing some clean performance.

In Table 10 we report the VTAB [Zhai et al., 2020] averaged performance over the categories *natural*, *specialized*, and *structured*. We observe that in clean evaluation, robust models sacrifice performance on *natural* and *specialized* (a trade-off between clean and robust performance is expected [Tsipras et al., 2019]). On *structured* the behavior is mixed - sometimes even outperforming the non-robust models. In the adversarial evaluation ( $\epsilon = 2/255$ ), we observe that the non-robust models are completely vulnerable, while our robust models maintain much better performance when attacked.

### D.2.2 Additional experiments on zero-shot text classification

In this section, we evaluate the zero-shot clean and adversarial accuracy of our models in additional text classification datasets. We follow the same attack setup as in the AG-News experiments, i.e.,

Table 9: **Zero-shot image classification.** We report the zero-shot image classification performance of original and bimodally robust models.

Model		Robust	CalTech101	Cars	Cifar10	Cifar100	DTD	EuroSAT	FGVC	Flowers	ImageNet-r	ImageNet-s	PCAM	Pets	STL10	Mean
Clean	CLIP-ViT-L/14	✗	82.1	77.5	95.2	68.2	55.7	63.4	28.4	79.4	86.5	48.9	53.0	93.9	98.8	71.6
		✓	81.1	71.6	92.2	68.9	44.9	28.7	24.6	69.7	83.3	47.0	59.9	91.9	98.1	66.3
	OpenCLIP-ViT-H/14	✗	84.4	92.2	97.5	82.8	68.7	72.5	42.4	80.2	88.4	56.1	54.9	95.1	98.1	77.9
		✓	83.8	89.8	93.3	69.7	61.1	34.4	35.8	73.4	85.7	52.9	50.4	94.0	97.2	70.9
	OpenCLIP-ViT-g/14	✗	84.3	92.1	97.7	84.0	68.8	65.6	36.4	78.1	88.2	55.5	55.6	95.2	98.2	76.9
		✓	83.1	88.4	91.7	67.3	58.1	29.0	30.7	71.2	84.9	52.0	52.5	92.5	96.2	69.0
$\epsilon = 2/255$	CLIP-ViT-L/14	✗	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		✓	70.5	27.8	65.6	34.2	25.3	11.6	6.0	33.8	55.5	26.4	22.1	69.0	89.7	41.3
	OpenCLIP-ViT-H/14	✗	0.0	0.0	0.3	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		✓	70.7	55.6	65.0	38.4	32.5	7.7	5.8	39.5	58.3	31.0	37.9	66.0	87.9	45.9
	OpenCLIP-ViT-g/14	✗	0.0	0.0	0.1	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		✓	71.3	52.1	62.6	34.0	28.5	4.7	4.0	34.2	53.3	28.6	26.5	57.5	84.7	41.7

Table 10: **VTAB zero-shot image classification.** We report the zero-shot image classification performance of original and bimodally robust models on VTAB Zhai et al. [2020].

Model		Robust	Natural	Specialized	Structured
Clean	ViT-L/14	✗	74.4	63.5	11.9
		✓	68.5	41.9	13.3
	ViT-H/14	✗	78.7	57.0	11.7
		✓	74.8	45.6	11.8
	ViT-g/14	✗	79.5	62.9	12.5
		✓	72.4	51.4	11.4
$\epsilon = 2/255$	ViT-L/14	✗	0.0	0.0	0.0
		✓	42.4	10.6	3.9
	ViT-H/14	✗	0.1	0.0	0.0
		✓	44.9	14.6	3.6
	ViT-g/14	✗	0.0	0.0	0.0
		✗	41.0	9.5	1.9

we employ Charmer-20 at  $k = 1$  without semantic constraints to evaluate the performance on SST-2 [Socher et al., 2013], IMDB [Maas et al., 2011] and Yelp [Yelp, 2015, Zhang et al., 2015].

In Fig. 9 we report the zero-shot adversarial accuracy already reported in Fig. 4, with the addition of SafeCLIP [Poppi et al., 2024]. SafeCLIP obtains a considerably lower clean and adversarial accuracy in comparison to the other CLIP variants.

In Table 11 we can observe that similarly to the AG-News results in Table 2, the models with robust text encoders achieve higher adversarial accuracy in the text domain, with improvements of more than 9.9 robust accuracy points for all models and datasets.

In Table 12, we present the clean and adversarial zero-shot accuracy when employing only the text encoder for the ViT-L/14 models. For that, we encode on sentence per label instead of one image per label as done in the main text. See Table 5 for more details on the sentences employed for the labels. We can observe that the adversarial accuracy is larger after adversarial finetuning with LEAF. Nevertheless, the clean and adversarial performance are worse when doing text-encoder-only zero-shot classification, e.g., a clean accuracy in AG-News with ViT-L/14 of 74.4 when using images as labels (Table 2) v.s. 54.8 when using sentences as labels.

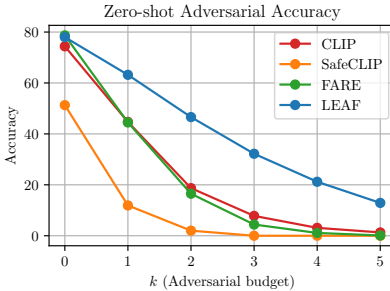


Figure 9: **Larger perturbations:** We evaluate the adversarial accuracy in AG-News for  $k \in \{1, 2, 3, 4, 5\}$  in the ViT-L/14 scale. Our model (LEAF) obtains the highest adversarial accuracy at all values of the distance bound  $k$ .

Table 11: **Zero-shot text classification.** We report the zero-shot text classification performance of original and bimodally robust models.

	Model	Robust	SST-2	IMDB	Yelp
Clean	CLIP-ViT-L/14	✗	71.2	61.6	80.9
		✓	71.9	61.4	82.0
	OpenCLIP-ViT-H/14	✗	61.6	57.5	73.7
		✓	58.4	53.2	72.6
$k=1$	OpenCLIP-ViT-g/14	✗	57.8	56.8	71.9
		✓	56.0	54.0	71.1
	CLIP-ViT-L/14	✗	6.8	13.7	21.0
		✓	23.2	31.0	43.8
$k=1$	OpenCLIP-ViT-H/14	✗	16.2	31.1	22.1
		✓	36.4	43.9	40.8
	OpenCLIP-ViT-g/14	✗	21.4	31.4	26.0
		✓	34.2	41.3	39.4

Table 12: **Text-encoder-only zero-shot text classification:** We report the clean and adversarial zero shot accuracy at  $k = 1$  employing only text-encoders. The adversarial accuracy improves after adversarial finetuning with LEAF. Nevertheless, employing only the text encoder provides worse clean and adversarial performance than employing images as labels as Qin et al. [2023].

Robust	AG-News		SST-2		IMDB		Yelp	
	Acc.	Adv.	Acc.	Adv.	Acc.	Adv.	Acc.	Adv.
✗	54.8	17.9	60.3	3.2	54.0	24.9	59.9	29.5
✓	53.5	34.7	58.9	24.1	51.5	44.9	56.7	47.5



### D.3 Additional experiments in text-to-image models

In this section, we provide additional experiments and examples for the text-to-image generation task. In Tables 13 and 14 we present the generation results in SD-1.5 and SDXL in the MS-COCO dataset and the first 5,000 images of the Flickr30k dataset. We measure the CLIPScore between the original caption and the generated image (T-I), the CLIPScore between the original image and the generated one (I-I), the attack objective (Eq. (2)) and for SD-1.5, the percentage of generated images triggering the NSFW filter (NSFW %). We can observe that the text encoders finetuned with LEAF, provide a higher generation quality for  $k > 1$  according to all generation metrics. Surprisingly, for  $k = 2$  and  $k = 4$  in the MS-COCO dataset, our text encoders triggered the NSFW filter less frequently than SafeCLIP [Poppi et al., 2024], which is specifically designed to avoid generating NSFW content.

In Tables 15 to 18 we present examples of the attacks on the first 10 samples of each dataset for both SD-1.5 and SDXL at  $k = 2$ . We can observe, that our text encoders provide qualitatively better images. The models with the original text encoders, provide images unrelated to the original image and caption more often than the models employing our text encoders.

In Table 19 we include the generation results with FLUX.1-dev [Black Forest Labs et al., 2025]. Since FLUX.1-dev employs CLIP ViT-L/14 and FLAN-T5 XXL [Chung et al., 2022] as text encoders, the model can only be benefited from our approach by replacing the CLIP text encoder with our LEAF counterpart. Similarly, we only attack the CLIP / LEAF text encoders and assume no access to FLAN-T5 XXL. Due to the high resolution of the FLUX.1-dev generations ( $1024 \times 1024$ ), we restrict the evaluation of FLUX.1-dev to the first 100 images in the MS-COCO validation set.

#### D.3.1 Transfer attacks on text-to-image models

In this section we evaluate the performance of transfer attacks on SD-1.5 with CLIP and LEAF as either the source model where the attack is optimized or the target model used for the image generation. In Table 20 we can observe that, as expected, when the source is equal to the target, the generated image quality is degraded the most. Our text encoder improves the generation quality in all cases except when the source is LEAF and  $k = 1$ , where CLIP obtains 0.04 more CLIPScore T2I score points than LEAF in this advantageous setup.

#### D.3.2 Preliminary study of typographic attacks

In this section we evaluate how our text encoder preserves the image quality under typographic prompts, i.e., prompts where characters have been changed for visually similar ones. To do so, we employ SD-1.5 and replace every “i” for a “1”, every “e” for a “3”, every “o” for a “0” and every “a” for an “@” in the first 100 prompts in the MS-COCO dataset. As an example, the first COCO caption turns into “A w0m@n st@nds ln th3 d1n lng @r3@ @t th3 t@bl3.”

In Table 21, we can observe that while the image generation quality with both encoders is quite low, using LEAF provides an improvement of 0.62 points in CLIPScore T2I and 2.77 in CLIPScore I2I.

### D.4 Embedding inversion examples

In Tables 22 and 23 we present examples from the embedding-to-text reconstructions results performed in Section 4.6.

### D.5 Additional retrieval experiments

For 1,000 validation set queries, the attack explained in the main part maximizes the similarity between the test query and a target string using different variants of the Charmer attack. In Table 24, we show the individual attack results across 3 target strings for differently trained LEAF models. One sees that on increasing training  $\rho$ , the robustness goes up with a slight decay in the clean retrieval performance. This trade-off is similar to the one seen for classification tasks in Fig. 3.

In Fig. 10, we visualize the top-3 retrieved images for the original and the perturbed queries. Although in some cases the non robust model retrieves a relevant query, the top-1 retrieved image is always different for clean and perturbed queries. However, the robust model always preserves the original top-1 retrieved image showing its robustness to such character perturbed queries.

Table 13: **Text-to-image generation results on MS-COCO:** SD-1.5 and SDXL are evaluated over the full 5000 images in the validation set. FLUX.1-dev is evaluated over the first 100 images due to the high resolution of the generated images.

Pipeline	k	Text encoder	$\text{Sim}(f_{\theta}(S), f_{\theta}(S'))$	CLIPScore T2I	CLIPScore I2I	NSFW (%)
SD-1.5	0	CLIP	-	<b>31.50</b> <sub>(±2.87)</sub>	<b>73.31</b> <sub>(±10.21)</sub>	0.64
		SafeCLIP	-	30.96 <sub>(±2.93)</sub>	73.27 <sub>(±10.08)</sub>	<b>0.44</b>
		LEAF	-	31.00 <sub>(±2.94)</sub>	73.06 <sub>(±10.12)</sub>	0.46
	1	CLIP	55.85 <sub>(±8.66)</sub>	27.53 <sub>(±4.52)</sub>	65.38 <sub>(±12.71)</sub>	0.96
		SafeCLIP	71.62 <sub>(±8.32)</sub>	27.43 <sub>(±4.09)</sub>	66.90 <sub>(±11.56)</sub>	<b>0.48</b>
		LEAF	<b>86.58</b> <sub>(±4.84)</sub>	<b>27.96</b> <sub>(±3.48)</sub>	<b>68.01</b> <sub>(±11.17)</sub>	0.50
	2	CLIP	33.18 <sub>(±9.29)</sub>	22.96 <sub>(±5.79)</sub>	57.21 <sub>(±13.90)</sub>	2.16
		SafeCLIP	50.87 <sub>(±10.34)</sub>	23.75 <sub>(±5.02)</sub>	61.02 <sub>(±12.06)</sub>	1.08
		LEAF	<b>73.15</b> <sub>(±7.45)</sub>	<b>25.23</b> <sub>(±4.36)</sub>	<b>63.40</b> <sub>(±11.95)</sub>	<b>0.62</b>
	3	CLIP	20.38 <sub>(±8.93)</sub>	19.45 <sub>(±5.86)</sub>	51.55 <sub>(±13.40)</sub>	2.52
		SafeCLIP	35.93 <sub>(±11.06)</sub>	20.41 <sub>(±5.61)</sub>	55.98 <sub>(±12.07)</sub>	<b>1.10</b>
		LEAF	<b>60.00</b> <sub>(±9.07)</sub>	<b>22.59</b> <sub>(±5.16)</sub>	<b>59.02</b> <sub>(±12.19)</sub>	1.26
4	CLIP	12.83 <sub>(±8.80)</sub>	17.42 <sub>(±5.68)</sub>	48.34 <sub>(±12.66)</sub>	2.70	
	SafeCLIP	26.05 <sub>(±11.04)</sub>	17.94 <sub>(±5.57)</sub>	52.31 <sub>(±11.57)</sub>	1.56	
	LEAF	<b>49.35</b> <sub>(±9.55)</sub>	<b>20.25</b> <sub>(±5.44)</sub>	<b>55.36</b> <sub>(±12.33)</sub>	<b>1.44</b>	
SDXL	0	CLIP + OpenCLIP	-	<b>31.90</b> <sub>(±2.84)</sub>	<b>71.87</b> <sub>(±10.58)</sub>	-
		2×LEAF	-	31.80 <sub>(±2.86)</sub>	71.78 <sub>(±10.60)</sub>	-
	1	CLIP + OpenCLIP	67.65 <sub>(±7.46)</sub>	28.33 <sub>(±4.11)</sub>	64.45 <sub>(±12.25)</sub>	-
		2×LEAF	<b>88.15</b> <sub>(±4.44)</sub>	<b>29.37</b> <sub>(±3.46)</sub>	<b>67.25</b> <sub>(±11.54)</sub>	-
	2	CLIP + OpenCLIP	47.58 <sub>(±8.74)</sub>	24.65 <sub>(±5.25)</sub>	57.97 <sub>(±12.89)</sub>	-
		2×LEAF	<b>76.49</b> <sub>(±7.12)</sub>	<b>27.14</b> <sub>(±4.33)</sub>	<b>63.27</b> <sub>(±12.19)</sub>	-
	3	CLIP + OpenCLIP	34.22 <sub>(±8.90)</sub>	21.45 <sub>(±5.70)</sub>	53.37 <sub>(±12.78)</sub>	-
		2×LEAF	<b>64.62</b> <sub>(±9.24)</sub>	<b>24.69</b> <sub>(±5.16)</sub>	<b>59.38</b> <sub>(±12.66)</sub>	-
4	CLIP + OpenCLIP	25.93 <sub>(±8.74)</sub>	19.07 <sub>(±5.60)</sub>	49.92 <sub>(±12.21)</sub>	-	
	2×LEAF	<b>54.08</b> <sub>(±10.22)</sub>	<b>22.45</b> <sub>(±5.67)</sub>	<b>55.70</b> <sub>(±12.85)</sub>	-	
FLUX.1-dev	0	CLIP + FLAN-T5 XXL	-	<b>30.56</b> <sub>(±2.86)</sub>	<b>71.19</b> <sub>(±12.13)</sub>	-
		LEAF + FLAN-T5 XXL	-	30.55 <sub>(±2.90)</sub>	71.18 <sub>(±12.83)</sub>	-
	1	CLIP + FLAN-T5 XXL	57.86 <sub>(±8.70)</sub>	<b>29.14</b> <sub>(±3.76)</sub>	68.09 <sub>(±12.82)</sub>	-
		LEAF + FLAN-T5 XXL	<b>87.07</b> <sub>(±4.52)</sub>	28.90 <sub>(±3.60)</sub>	<b>68.79</b> <sub>(±12.91)</sub>	-
	2	CLIP + FLAN-T5 XXL	35.04 <sub>(±8.87)</sub>	27.03 <sub>(±5.20)</sub>	63.60 <sub>(±13.51)</sub>	-
		LEAF + FLAN-T5 XXL	<b>73.70</b> <sub>(±6.90)</sub>	<b>27.38</b> <sub>(±4.09)</sub>	<b>65.66</b> <sub>(±13.01)</sub>	-
	3	CLIP + FLAN-T5 XXL	21.84 <sub>(±7.78)</sub>	24.47 <sub>(±6.00)</sub>	59.40 <sub>(±14.09)</sub>	-
		LEAF + FLAN-T5 XXL	<b>59.83</b> <sub>(±9.23)</sub>	<b>25.71</b> <sub>(±5.16)</sub>	<b>62.11</b> <sub>(±13.84)</sub>	-
4	CLIP + FLAN-T5 XXL	14.79 <sub>(±7.10)</sub>	22.72 <sub>(±6.11)</sub>	57.68 <sub>(±14.33)</sub>	-	
	LEAF + FLAN-T5 XXL	<b>49.57</b> <sub>(±9.86)</sub>	<b>23.51</b> <sub>(±5.98)</sub>	<b>59.59</b> <sub>(±15.27)</sub>	-	

Table 14: Text-to-image generation results on Flickr30k:

Pipeline	k	Text encoder	Sim( $f_{\theta}(S), f_{\theta}(S')$ )	CLIPScore T2I	CLIPScore I2I	NSFW (%)
SD-1.5	0	CLIP	-	<b>33.27</b> ( $\pm 3.21$ )	<b>71.27</b> ( $\pm 10.20$ )	0.42
		SafeCLIP	-	32.16( $\pm 3.35$ )	70.20( $\pm 10.25$ )	0.42
		LEAF	-	32.63( $\pm 3.17$ )	70.73( $\pm 10.23$ )	<b>0.26</b>
	1	CLIP	63.48( $\pm 9.01$ )	<b>30.72</b> ( $\pm 4.16$ )	66.43( $\pm 11.25$ )	0.84
		SafeCLIP	77.31( $\pm 7.11$ )	29.32( $\pm 4.19$ )	65.68( $\pm 10.85$ )	0.92
		LEAF	<b>89.80</b> ( $\pm 3.89$ )	30.37( $\pm 3.56$ )	<b>67.54</b> ( $\pm 10.56$ )	<b>0.66</b>
	2	CLIP	42.37( $\pm 10.21$ )	27.71( $\pm 5.18$ )	61.28( $\pm 12.18$ )	1.28
		SafeCLIP	59.79( $\pm 9.63$ )	26.24( $\pm 4.72$ )	61.66( $\pm 11.12$ )	0.87
		LEAF	<b>79.28</b> ( $\pm 6.55$ )	<b>28.43</b> ( $\pm 4.05$ )	<b>64.66</b> ( $\pm 10.80$ )	<b>0.68</b>
SDXL	0	CLIP + OpenCLIP	-	<b>33.85</b> ( $\pm 3.24$ )	<b>69.07</b> ( $\pm 10.54$ )	-
		2×LEAF	-	33.82( $\pm 3.22$ )	69.06( $\pm 10.50$ )	-
	1	CLIP + OpenCLIP	75.15( $\pm 6.33$ )	31.24( $\pm 4.00$ )	64.03( $\pm 11.23$ )	-
		2×LEAF	<b>91.32</b> ( $\pm 3.40$ )	<b>31.63</b> ( $\pm 3.54$ )	<b>65.87</b> ( $\pm 10.89$ )	-
	2	CLIP + OpenCLIP	58.02( $\pm 8.49$ )	28.30( $\pm 4.81$ )	59.09( $\pm 11.47$ )	-
		2×LEAF	<b>82.82</b> ( $\pm 5.84$ )	<b>29.83</b> ( $\pm 4.09$ )	<b>63.03</b> ( $\pm 11.15$ )	-

Table 15: Attack examples on MS-COCO with SD-1.5 at  $k = 2$ : The color borders indicate null, partial and total matching to the original image caption. The model with the original text encoder provides images involving a footballer, a lizard or a gun, when prompted about a bear, a women skiing or a group of people respectively. With our text encoders, the generation does not drift in topic so much.

ID	Original caption	Original image	Original		SafeCLIP		LEAF	
			Adversarial caption	Generated image	Adversarial caption	Generated image	Adversarial caption	Generated image
139	A woman stands in the dining area at the table.		A woman stands in the dining area at the table-		A woman stands in the dining area at the table.		A woman stands in the dining area at the table.	
285	A big burly grizzly bear is show with grass in the background.		A big burly grizzly bear is show with g'rass in the background.		A big burly grizzly bear is show with grass in the background.		A big burly grizzly bear is show with @rass in the background.	
632	Bedroom scene with a bookcase, blue comforter and window.		Bedroom scene with a bookcase, blue comfor#ter and window.		Bedroom scene with a @ookcase, bl#ue comforter and window.		Bedroom scene with a kookcase, blue comforter and window.	
724	A stop sign is mounted upside-down on it's post.		A stop sign is mounted upside-down on it's post.		A stop sign is mounted upside-down on it's pos\$.		A stop sign is mounted upside-down on it's post.	
776	Three teddy bears, each a different color, snuggling together.		Thre#e teddy bears, each a different color, snuggling together.		Three teddy bears, eac= a different color, snuggling together.		9hree teddy bears, each a different color, snuggling together.	
785	A woman posing for the camera standing on skis.		A woma#n posing for the camera standing on >kis.		A woma6 posing for the camera standing onuskis.		A -oman posing for the camera standing onoskis.	
802	A kitchen with a refrigerator, stove and oven with cabinets.		A kit>chen with a refrigerator, stove and oven withmcabinets.		A kilchen with a refr#igerator, stove and oven with cabinets.		Aqkitchen withra refrigerator, stove and oven with cabinets.	
872	A couple of baseball player standing on a field.		A couple of bas#ball player standing on a fit#eld.		A cozuple of basbal#m player standing on a field.		A coupl. of baseball player standing on a 'ield.	
885	a male tennis player in white shorts is playing tennis		a male ten#is player in white shor'ts is playing tennis		a male tennis player in wh.ite h#orts is playing tennis		a male tennis player in white shorts is playing tennis	
1000	The people are posing for a group photo.		The pzople are posing for a group ph#oto.		The people are posing for a gr1oup photo.		The people are posing forza group photo.	

Table 16: Attack examples on MS-COCO with SDXL at  $k = 2$ :

ID	Original caption	Original image	Original		LEAF	
			Adversarial caption	Generated image	Adversarial caption	Generated image
139	A woman stands in the dining area at the table.		A woma8 stands in the jining area at the table.		3 woman' stands in the dining area at the table.	
285	A big burly grizzly bear is show with grass in the background.		A big burly grlizzly bear is show with @rass in the background.		A big burly !rizzly bear is show with krass in the background.	
632	Bedroom scene with a bookcase, blue comforter and window.		Bedroom scene with a zookcase, blue comforter and window.		Bedroom scene with a cookcase, blue cosmforter and window.	
724	A stop sign is mounted upside-down on it's post.		A stop gign is mounted pupside-down on it's post.		A 3top sign is mounted upside-downnton it's post.	
776	Three teddy bears, each a different color, snuggling together.		Thre:e teddy bears, each a different color, snuggling toge—ther.		ahree teddy bears, each a different color, snuggling toge,ther.	
785	A woman posing for the camera standing on skis.		A woma: posing for the camera standing ontskis.		A -oman posing for the camera standing onoskis.	
802	A kitchen with a refrigerator, stove and oven with cabinets.		A ki:chen with a refr@igerator, stove and oven with cabinets.		Aqkitchen withra refrigerator, stove and oven with cabinets.	
872	A couple of baseball player standing on a field.		A couple of basebill player standing on a #ield.		A coupl of baseball player standing on a qield.	
885	a male tennis player in white shorts is playing tennis		a male tennis pl*ayer in white #horts is playing tennis		aemale tennis playerein white shorts is playing tennis	
1000	The people are posing for a group photo.		The neople are posing for a group  hoto.		The peoplecare posing forza group photo.	

### D.5.1 Bimodal attacks in text-to-image retrieval

Building on top of text-modality robustness for text-to-image retrieval from the main part, we now assess the robustness to bimodal attacks for both the image and text modalities for  $1k$  samples of the MS-COCO test set. The evaluation starts from the known baseline ( $k = 1$  text perturbations) from Table 3 and applies an untargeted adversarial attack to the images. We use APGD [Croce and Hein, 2020] for 100 iterations with small  $\ell_\infty$  perturbation radii of  $2/255$  and  $4/255$ . This perturbation is designed to maximize the distance between the original and perturbed image embeddings, thereby disrupting the model’s ability to retrieve the correct text. This attack protocol, is similar to CoAttack [Zhang et al., 2022], where the text attack follows the image attack.

The results in Table 25 highlight the superior resilience of the LEAF-trained models. For the critical recall@1 metric, LEAF improved retrieval performance by nearly 7% over the baseline across both perturbation radii. Importantly, this significant gain in robustness did not come at the cost of clean performance (performance on clean data), as indicated by the ‘clean’ column results. This finding strongly underscores the importance of dual modality robustness: the ability to maintain high performance despite adversarial attacks on either the image or text data, making LEAF the most robust solution in this challenging setup.

Table 17: Attack examples on Flickr30k with SD-1.5 at  $k = 2$ :

ID	Original caption	Original image	Original		SafeCLIP		LEAF	
			Adversarial caption	Generated image	Adversarial caption	Generated image	Adversarial caption	Generated image
1000092795	Two young guys with shaggy hair look at their hands while hanging out in the yard .		Two young guys with shaggy hair look at their hands while hanging out in the yard .		Two young guys with shaggy hair look at their hands while hanging out in the yard .		Two young guys with shaggy hair look at their hands while hanging out in the yard .	
10002456	Several men in hard hats are operating a giant pulley system .		Several men in hard hats are operating a giant pulley system .		Several men in hard hats are operating a giant pulley system .		Several men in hard hats are operating a giant pulley system .	
1000268201	A child in a pink dress is climbing up a set of stairs in an entry way .		A child in a pink dress is climbing up a set of stairs in an entry way .		A child in a pink dress is climbing up a set of stairs in an entry way .		A child in a pink dress is climbing up a set of stairs in an entry way .	
1000344755	Someone in a blue shirt and hat is standing on stair and leaning against a window .		Someone in a blue shirt and hat is standing on stair and leaning against a window .		Someone in a blue shirt and hat is standing on stair and leaning against a window .		Someone in a blue shirt and hat is standing on stair and leaning against a window .	
1000366164	Two men , one in a gray shirt , one in a black shirt , standing near a stove .		Two men , one in a gray shirt , one in a black shirt , standing near a stove .		Two men , one in a gray shirt , one in a black shirt , standing near a stove .		Two men , one in a gray shirt , one in a black shirt , standing near a stove .	
1000523639	Two people in the photo are playing the guitar and the other is poking at him .		Two people in the photo are playing the guitar and the other is poking at him .		Two people in the photo are playing the guitar and the other is poking at him .		Two people in the photo are playing the guitar and the other is poking at him .	
1000919630	A man sits in a chair while holding a large stuffed animal of a lion .		A man sits in a chair while holding a large stuffed animal of a lion .		A man sits in a chair while holding a large stuffed animal of a lion .		A man sits in a chair while holding a large stuffed animal of a lion .	
10010052	A girl is on rollerskates talking on her cellphone standing in a parking lot .		A girl is on rollerskates talking on her cellphone standing in a parking lot .		A girl is on rollerskates talking on her cellphone standing in a parking lot .		A girl is on rollerskates talking on her cellphone standing in a parking lot .	
1001465944	An asian man wearing a black suit stands near a dark-haired woman and a brown-haired woman .		An asian man wearing a black suit stands near a dark-haired woman and a brown-haired woman .		An asian man wearing a black suit stands near a dark-haired woman and a brown-haired woman .		An asian man wearing a black suit stands near a dark-haired woman and a brown-haired woman .	
1001545525	Two men in Germany jumping over a rail at the same time without shirts .		Two men in Germany jumping over a rail at the same time without shirts .		Two men in Germany jumping over a rail at the same time without shirts .		Two men in Germany jumping over a rail at the same time without shirts .	

Table 18: Attack examples on Flickr30k with SDXL at  $k = 2$ :

ID	Original caption	Original image	Original		LEAF	
			Adversarial caption	Generated image	Adversarial caption	Generated image
1000092795	Two young guys with shaggy hair look at their hands while hanging out in the yard .		Two young guys with shaggyhair look at their hands while hanging out in the  ard .		Two young guys with shaggychair look at their hands while hanging out in the  ard .	
10002456	Several men in hard hats are operating a giant pulley system .		Several men in \$ard hats are operating a gi-ant !ulley system .		Several men in hardchats are operating a giant sulley system .	
1000268201	A child in a pink dress is climbing up a set of stairs in an entry way .		A ch ld in a pink dr..ss is climbing up a set of stairs in an entry way .		A chwild in a pink dress is climbing up a set of stairs in an entry way .	
1000344755	Someone in a blue shirt and hat is standing on stair and leaning against a window .		Someone in a bl2ue shirt and hat is standing on stair and leaning against a :indow .		Someone in a blue shirt andwhat is standing on &tair and leaning against a window .	
1000366164	Two men , one in a gray shirt , one in a black shirt , standing near a stove .		Twomen , one in a gray shirt , one in a black shirt , standing near a @tove .		Twomen , one in a gray shirt , one in a black shirt , standing near a ptove .	
1000523639	Two people in the photo are playing the guitar and the other is poking at him .		Two people in the ph?oto are playing the gu#itar and the other is poking at him .		Two people in the photo are playing the suitarm and the other is poking at him .	
1000919630	A man sits in a chair while holding a large stuffed animal of a lion .		A man sits in a ch5ir while holding a large stu!ffed animal of a lion .		A man sits in a chair while holding a large stuffe. animal of aklion .	
10010052	A girl is on rollerskates talking on her cellphone standing in a parking lot .		A girl is on rollerskates talking on her cellphone standing in a parki{ng}lot .		A girl is on roller-skatesstalking on her cellphone standing in a parkingslot .	
1001465944	An asian man wearing a black suit stands near a dark-haired woman and a brown-haired woman .		Axn asian man wearing a black suit stands near a dark-haired woman #nd a brown-haired woman .		Axn asian man wearing a black suit stands near a dark-haired woman anz a brown-haired woman .	
1001545525	Two men in Germany jumping over a rail at the same time without shirts .		Two men in Germany jumping over a rai7 at the same time ?ithout shirts .		cwo men in Germany jumping over a rail at the same time ?ithout shirts .	

Table 19: Attack examples on MS-COCO with FLUX.1-dev at  $k = 2$ :

ID	Original caption	Original image	Original		LEAF	
			Adversarial caption	Generated image	Adversarial caption	Generated image
139	A woman stands in the dining area at the table.		A woma@ stands in the xining area at the table.		A vwoman stands in the dining area at the table.	
285	A big burly grizzly bear is show with grass in the background. Bedroom		A big burly griexly bear is show with ?rass in the background. Bedr=oom		A big burly .rizzly bear is show with @rass in the background. Bedroomascene	
632	scene with a bookcase, blue comforter and window.		scene with a bookcase, blue comfor#ter and window.		with a kookcase, blue comforter and window. A stopsign is	
724	A stop sign is mounted upside-down on it's post.		A stop \$ign is mounted upside-down on fit's post.		A stopsign is mounted upside-downnton it's post.	
776	Three teddy bears, each a different color, snuggling together.		+hree teddy bears, each a different color, snuggling @gether.		9hree teddy bears, each a different color, snuggling toge,ther.	
785	A woman posing for the camera standing on skis.		A woma7 posing for the camera standing onoskis.		A -oman posing for the camera standing onoskis.	
802	A kitchen with a refrigerator, stove and oven with cabinets.		A ki=chen with a refrigeratoa, stove and oven with cabinets.		Aqkitchen withra refrigerator, stove and oven with cabinets.	
872	A couple of baseball player standing on a field.		A couple of \$aseball player standing on a #field.		A coupl. of baseball player standing on a ^field.	
885	a male tennis player in white shorts is playing tennis		a malec tennis pl*ayer in white shorts is playing tennis		aimale tennis playerein white shorts is playing tennis	
1000	The people are posing for a group photo.		The neople are posing for a group ph?oto.		The people are posing forza group bhoto.	

Table 20: **Transfer attacks in SD-1.5**: Columns represent the source text encoder, where the attack is optimized, and rows the target text encoder, where the attack is evaluated. LEAF obtains the highest CLIPScores for every setup except the CLIPScore T2I at  $k = 1$  with LEAF as a source model.

k	Target \ Source	CLIPScore T2I		CLIPScore I2I	
		CLIP	LEAF	CLIP	LEAF
1	CLIP	27.53(±4.52)	<b>28.00</b> (±3.70)	65.38(±12.72)	66.73(±11.61)
	LEAF	<b>28.84</b> (±3.49)	27.96(±3.48)	<b>69.47</b> (±11.05)	<b>68.01</b> (±11.17)
2	CLIP	22.96(±5.80)	24.46(±4.86)	57.21(±13.90)	60.80(±12.52)
	LEAF	<b>26.72</b> (±4.23)	<b>25.23</b> (±4.36)	<b>66.11</b> (±11.88)	<b>63.40</b> (±11.95)
3	CLIP	19.45(±5.86)	21.30(±5.44)	51.55(±13.40)	55.68(±12.59)
	LEAF	<b>24.61</b> (±5.19)	<b>22.59</b> (±5.16)	<b>62.68</b> (±12.57)	<b>59.02</b> (±12.19)
4	CLIP	17.42(±5.68)	19.10(±5.48)	48.34(±12.66)	52.24(±12.20)
	LEAF	<b>22.44</b> (±5.78)	<b>20.25</b> (±5.44)	<b>59.25</b> (±12.95)	<b>55.36</b> (±12.33)

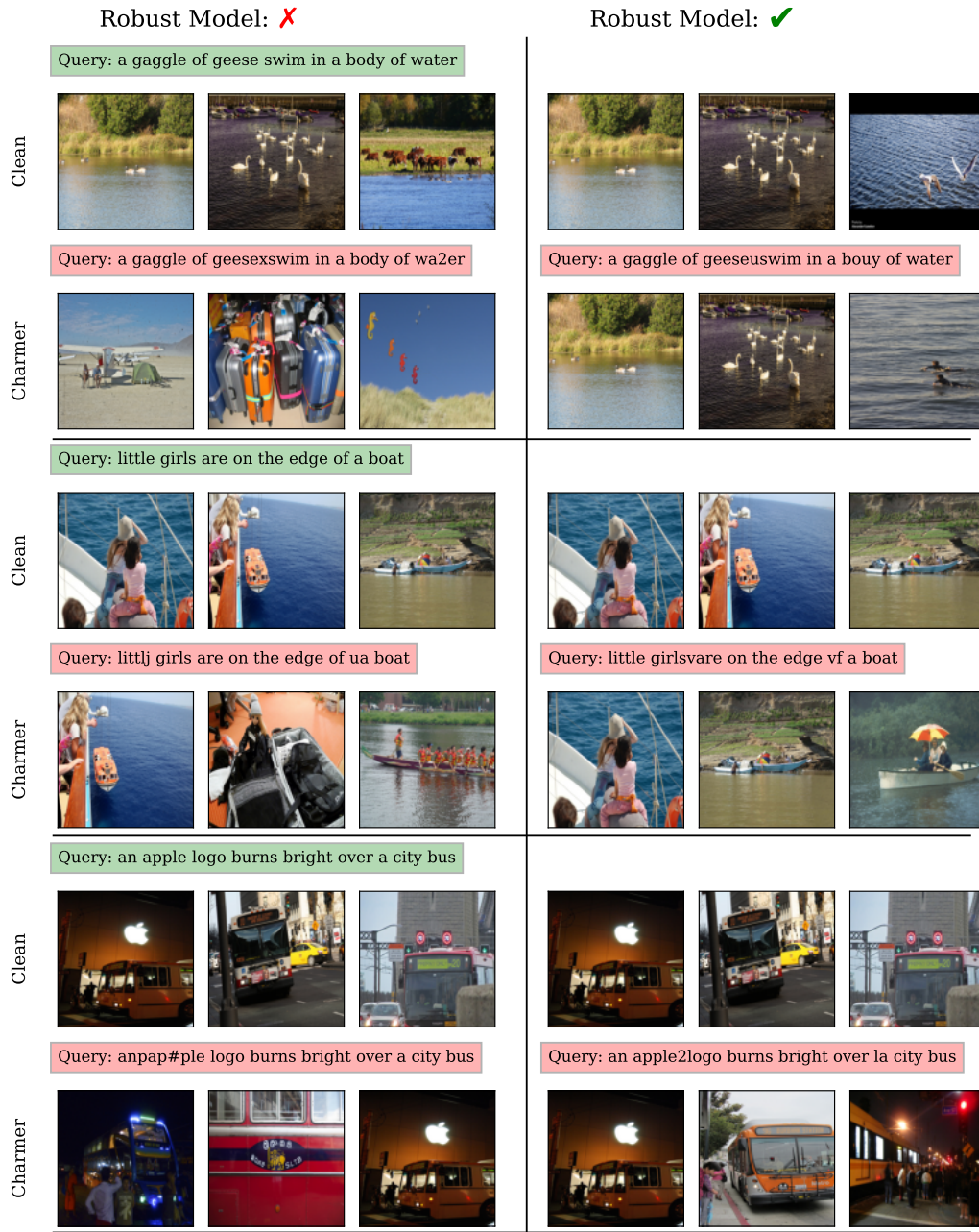


Figure 10: **Visualizing MS-COCO retrieved images.** For our ViT-L/14 robust model and its non-robust counterpart, we show the top-3 retrieved images for the original **Query** and the perturbed **Query** via the constrained Charmer ( $k = 2, n = 10$ ) attack. On average, the robust model is able to preserve the order and retrieves semantically relevant images (esp. top-1) even under perturbation.



Table 21: **Performance of SD-1.5 under typographic attacks:** The generation quality is low with both the original CLIP text encoder and the LEAF counterpart. As a reference, the generation quality of SD-1.5 with unperturbed inputs is a CLIPScore of 31.50 T2I and 73.31 I2I. However, LEAF is able to attain a higher score both in T2I and I2I CLIPScore.

Text encoder	CLIPScore T2I	CLIPScore I2I
CLIP	16.79( $\pm 4.63$ )	45.27( $\pm 13.12$ )
LEAF	<b>17.41</b> ( $\pm 4.27$ )	<b>48.04</b> ( $\pm 13.25$ )

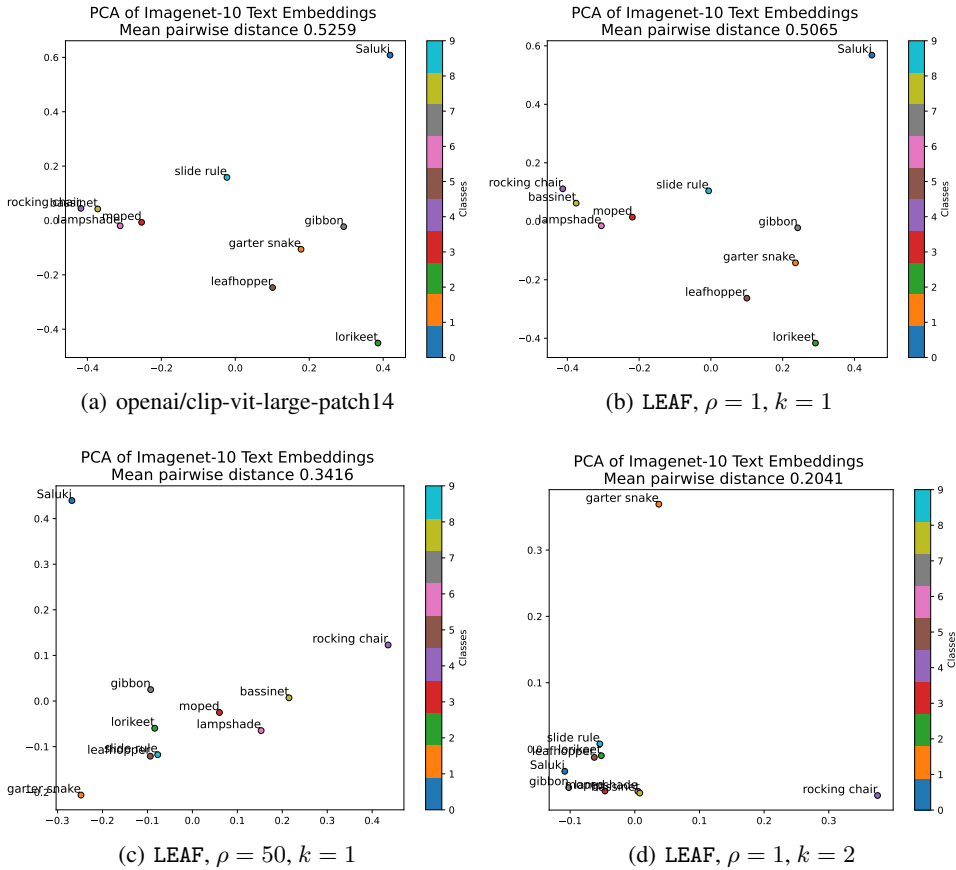


Figure 11: Ablation study on the cause of the clean performance drop in zero-shot classification.

## D.6 Ablation studies

In Appendix D.6.1 we evaluate the performance drop in zero-shot image classification when training without semantic constraints. In Appendix D.6.2 we measure the Eq. (TextFARE) loss before and after training.

### D.6.1 On the performance drop without semantic constraints

First, we perform an ablation study to better understand the cause of the performance drop in terms of clean accuracy in Table 8. We select 10 classes from the ImageNet dataset and visualize the corresponding text embeddings using the prompt “a photo of a LABEL”. In Fig. 11, we observe that as  $\rho$  and  $k$  increase, the class projections in 2D space become more clustered. We compute the mean pairwise distance, defined as the average L2 distance between all class pairs, and find that it decreases significantly.

Table 22: **Text embedding inversion examples for ViT-H/14.** We highlight in **red** words that are reconstructed by the robust model but not by the clean model; in **teal** words that are reconstructed by the clean model but not by the robust model; and in **yellow** words that are not reconstructed by either model. The robust model clearly misses fewer words.

Original	Robust	Reconstructed ViT-H/14
A car <b>and</b> a <b>public</b> transit vehicle <b>on</b> a road.	✗ ✓	public transit car alongside a vehicle amongst partially road road ." jrnnotified car and transit vehicle sit on a road ).
An <b>image</b> of a <b>hotel</b> bathroom that is ugly.	✗ ✓	ugly bathroom demonstrating poorly gross envir[U+0442]khobbhutto? ugly hotel bathroom showcasing concerns resemble ?magbbhutto.
An <b>older</b> picture of a large kitchen with <b>white</b> appliances.	✗ ✓	older earliest appenhistorical archival picture featuring older smaller large kitchen large kitchen pictured prior a a looked white appliances unidenti).
A girl sitting on a bench in front of a stone wall.	✗ ✓	prepped amina ssels sitting sitting bench near stone textured wall [U+1F91F]girl girl laghateparth girl twitart bench sitting outside a stone wall ??>."
A clean kitchen with the windows <b>white</b> and open.	✗ ✓	behold beautiful windows bein somewhere '?; white-beautifully clean kitchen a a kitchen with windows white wit yet clean .
Two women waiting at a bench next to a street.	✗ ✓	;/ /'": ;/ ,' two women waiting bench against street : two women waiting at an street bench ?bbcone .
An <b>office</b> cubicle with four different types of computers.	✗ ✓	four various computically cubicè compu?their desktop desk parked office cubic??eczw with four different computers either
An <b>old</b> victorian style bed frame in a bedroom.	✗ ✓	old ornate victorian bed showcasing ?wouldfeeold ). victorian finornate bed frame placed in a bedroom .
A <b>striped</b> plane flying up into the sky as the sun shines behind it.	✗ ✓	a sized ;/ wildly crafted plane near dramatically dramatically sun sunlight stripes approaching upward underneath a striped ??ûp plane coming above into sun ?[U+0648]sky .
A cat in between two cars in a parking lot.	✗ ✓	seemingly domestic cat sits standing among two cars in parking %. cat between two ?four cars docked paved parking lot .

Table 23: **Text embedding inversion examples for ViT-g/14.** We highlight reconstructions, we highlight in **red** words that are reconstructed by the robust model but not by the clean model; in **teal** words that are reconstructed by the clean model but not by the robust model; and in **yellow** words that are not reconstructed by either model. The robust model clearly misses fewer words.

Original	Robust	Reconstructed ViT-H/14
A car and a public transit vehicle on a road.	✗ ✓	partially tionally car sits alongside alongside roads public transit vehicle '. a car and eachother and a roadway public transit vehicle .
An image of a hotel bathroom that is ugly.	✗ ✓	apparent nicely tered hotel bathroom containing looking ugly pfmage image of a ugly と繋?* an hotel bathroom .
An older picture of a large kitchen with white appliances.	✗ ✓	a large kitchen photographed before that wasn resembled older . large old whil, an kitchen featuring reaswhite appliances
A girl sitting on a bench in front of a stone wall.	✗ ✓	girl near stone wall in a bench aciantly sitting tedly tedly ). girl sitting while a stone wall sits alongside an bench a jend_of_text <sub>i</sub> ).
A clean kitchen with the windows white and open.	✗ ✓	view of a white kitchen and nicely clean windows . an clean and white kitchen with windows thwindows .
Two women waiting at a bench next to a street.	✗ ✓	along a street bench . two women crouwaited stares . :// ; two women wait a street while bench outside .
An office cubicle with four different types of computers.	✗ ✓	office cubicle depicting four various different computers alongside payoff ). office cubicle containing an workplace with four different types computers
An old victorian style bed frame in a bedroom.	✗ ✓	eighsundaymotivation throwback© ?shutterintimacy "; victorian bed a victorian style bed frame uas in a bedroom .
A striped plane flying up into the sky as the sun shines behind it.	✗ ✓	nearly seemingly seemingly oooooooo a striped ambitious plane being flying into sky with sun light a striped plane being flying over above , but shining sun enguliot ung behind
A cat in between two cars in a parking lot.	✗ ✓	cat sitting through parked parking lot ?) alongside two two cars cat sits in an parking lot between two cars either ).

Table 24: **Detailed retrieval results for  $k = 2, n = 10$  constrained attack.** This is an extension of Table 3 for the ViT-L/14 model. We show how the robustness changes with changing training  $\rho$  across the three target texts.

Model	Train $\rho$	MS-COCO T→I retrieval			
		Clean		Charmer-Con	
		R@1	R@5	R@1	R@5
Target: A man aggressively kicks a stray dog on the street.					
non-robust	-	49.11	73.79	28.88	52.58
CLIP-ViT-L/14	1	49.33	73.98	37.34	62.16
CLIP-ViT-L/14	2	49.35	73.73	37.78	62.84
CLIP-ViT-L/14	5	49.63	73.82	38.66	63.86
CLIP-ViT-L/14	10	48.99	73.60	40.22	65.30
CLIP-ViT-L/14	20	48.97	73.72	37.92	62.44
CLIP-ViT-L/14	50	48.71	73.72	40.70	66.20
Target: This is an image of a a pyramid.					
non-robust	-	49.11	73.79	31.90	54.90
CLIP-ViT-L/14	1	49.33	73.98	36.30	60.08
CLIP-ViT-L/14	2	49.35	73.73	39.55	64.65
CLIP-ViT-L/14	5	49.63	73.82	40.38	65.34
CLIP-ViT-L/14	10	48.99	73.60	37.60	62.20
CLIP-ViT-L/14	20	48.97	73.72	40.00	65.46
CLIP-ViT-L/14	50	48.71	73.72	41.42	66.66
Target: A group of teenagers vandalizes a public statue.					
non-robust	-	49.11	73.79	30.68	54.22
CLIP-ViT-L/14	1	49.33	73.98	35.26	59.36
CLIP-ViT-L/14	2	49.35	73.73	39.29	63.76
CLIP-ViT-L/14	5	49.63	73.82	36.74	61.36
CLIP-ViT-L/14	10	48.99	73.60	41.42	65.50
CLIP-ViT-L/14	20	48.97	73.72	41.04	66.12
CLIP-ViT-L/14	50	48.71	73.72	38.56	62.38

Table 25: **Bimodal attacks in MS-COCO text-to-image retrieval.** Following [Zhang et al., 2022], we attack the vision-only robust (FARE) and our bimodally robust LEAF models. First we attack the text modality with Charmer-Con ( $k = 1$ ) and then use APGD with 100 iterations to perturb input images.

Method	Recall@1			Recall@5		
	Clean	$\epsilon = \frac{2}{255}, k = 1$	$\epsilon = \frac{4}{255}, k = 1$	Clean	$\epsilon = \frac{2}{255}, k = 1$	$\epsilon = \frac{4}{255}, k = 1$
Original	48.9	17.2	8.9	73.1	35.2	19.7
FARE	49.1	36.6	35.8	73.8	62.2	61.0
LEAF	48.7	43.4	42.8	73.7	67.4	66.9

### D.6.2 On the Eq. (TextFARE) loss

In this section, we evaluate the effectiveness of our method LEAF in minimizing the loss in Eq. (TextFARE). First, we measure the loss before and after adversarial finetuning in the ViT-L/14 scale on the first 100 images in the AG-News dataset at  $k = 1$ . We evaluate the inner max of Eq. (TextFARE) with the LEAF attack with and without semantic constraints (Appendix D.1) and with  $\rho \in \{1, 2, 5, 10, 20, 50\}$ . As baselines, we evaluate the same term with the Charmer-20 attack and a Bruteforce approach, which evaluates all of the possible sentences at Levenshtein distance  $k = 1$ .

In Fig. 12 we can observe that training with LEAF, we generalize to be robust to stronger attacks, even if they do not employ semantic constraints. For all cases, employing a larger  $\rho$  reduces the

Table 26: **Evaluating the loss in Eq. (FARE) and Eq. (TextFARE) across different scales:** We evaluate the ViT-L/14, ViT-H/14 and ViT-g/14 with and without our adversarial finetuning (LEAF) in both the image (ImageNet) and text domain (AG-News).  $L_{\text{clean}}$  refers to the respective loss when there is no perturbation applied, thus measuring the deviation to the original model. Robust models present a lower adversarial loss in both domains, with larger models presenting a higher loss before and after adversarial finetuning due to the use of larger embedding dimensions.

Model	Robust	ImageNet		AG-News		
		$L_{\text{clean}}$	$L_{\text{adv}}$	$L_{\text{clean}}$	$L_{\text{adv-cons.}}$	$L_{\text{adv-uncons.}}$
ViT-L/14	✗	0.0	789.7	0.0	58.4	82.6
ViT-L/14	✓	33.1	56.4	6.8	23.6	41.7
ViT-H/14	✗	0.0	1042.8	0.0	73.4	111.3
ViT-H/14	✓	47.9	89.6	13.3	40.7	76.3
ViT-g/14	✗	0.0	2172.5	0.0	112.3	175.0
ViT-g/14	✓	93.6	181.2	18.8	66.0	121.6

gap between the LEAF estimate and the true inner max of Eq. (TextFARE), i.e., Bruteforce. After adversarial finetuning with LEAF, both the loss estimates with Charmer-20 and Bruteforce are reduced.

Then, we evaluate the inner max of Eq. (TextFARE) in the ViT-L/14, ViT-H/14 and ViT-g/14 scales with Charmer-20 before and after adversarial finetuning with LEAF. Similarly, the Charmer-20 loss is minimized even if no semantic constraints are used in the estimate, for all model sizes. The loss is larger for larger model sizes both before and adversarial finetuning. This could be due to the larger embedding dimension for the ViT-H/14 and ViT-g/14 models. Finally, we also evaluate the inner max of Eq. (FARE) in the image domain. To this end, we compute adversarial perturbations for 100 ImageNet images with a 100-steps APGD attack on the Eq. (FARE) objective at radius  $\epsilon = 2/255$ . The results are reported in Table 26: similar to the textual attacks, we observe that the loss increases with model size. Importantly, the robust models generally demonstrate much smaller adversarial loss than their original counterparts. These results validate the intuition from Fig. 1 (left): the robust models map perturbed inputs much closer to the original inputs than the original models.

### D.6.3 Performance under token-level attacks

In this section, we evaluate the performance of our LEAF ViT-L/14, ViT-H/14 and ViT-g/14 models under the TextFooler token-level adversarial attack [Jin et al., 2020]. Furthermore, we replicate the experiment by Abad Rocamora et al. [2024] and finetune BERT-base [Devlin et al., 2019] on the SST-2 dataset with our character-level attack to evaluate the character-level and token-level accuracy of the classifier.

Abad Rocamora et al. [2024] conclude that token-level defenses are not effective for character-level attacks and vice-versa. In Table 28, we can observe that the in line with their results, character-level defenses are not effective for the token-level TextFooler attack.

In Table 27 we present the BERT-base Adversarial Training results. In line with the results of [Abad Rocamora et al., 2024], we observe that adversarial training with character-level attacks does not improve the robustness in the token level. Regarding character-level robustness, we observe that LEAF obtains almost 5 points less in adversarial accuracy with respect to training with Charmer, but preserves a clean accuracy 4 points higher.

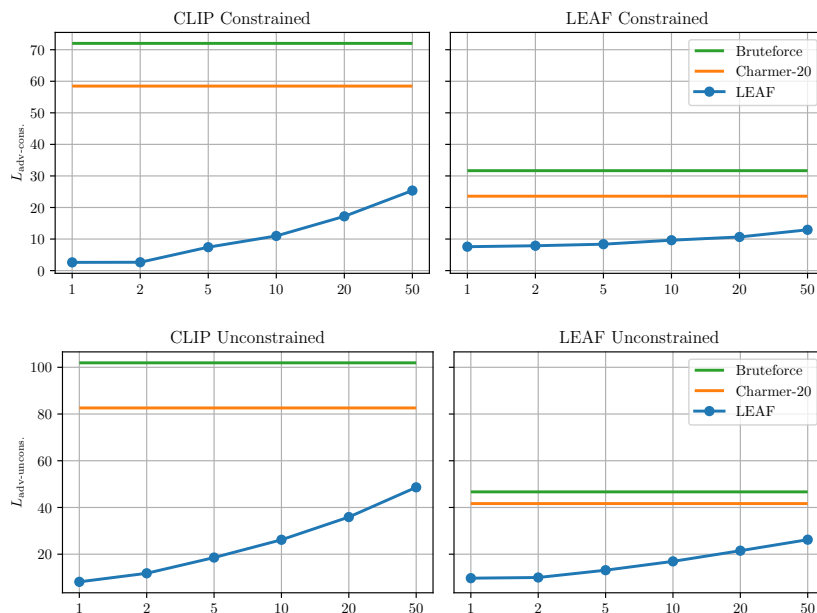


Figure 12: **Evaluating the loss in Eq. (TextFARE) with different attacks:** We evaluate the models in the ViT-L/14 scale on the first 100 sentences in the AG-News test dataset. For increasing values of  $\rho$ , the LEAF attack approximates better the inner max in Eq. (TextFARE), getting closer to the Bruteforce maximum. Our models, trained with LEAF and  $\rho = 50$ , reduce the Bruteforce loss, meaning that our models generalize to stronger attacks.

Table 27: **Adversarial Training of BERT-base models in SST-2:** We report the clean accuracy, character-level (Charmer) adversarial accuracy and token-level (TextFooler) adversarial accuracy.

Method	Acc.	Adv. (Charmer)	Adv. (TextFooler)
Original*	<b>92.43</b>	33.26	4.47
TextGrad* [Hou et al., 2023]	80.94	26.44	<b>23.18</b>
Charmer* [Abad Rocamora et al., 2024]	87.20	<b>69.46</b>	4.21
LEAF	91.51	64.68	5.50
LEAF-constrained	91.86	62.27	4.13

\* Numbers from Abad Rocamora et al. [2024]. The results were obtained as an average of 5 training runs.

Table 28: **Token-level adversarial attacks in zero-shot text classification.** We report the TextFooler adversarial accuracy (Adv.) on on AG-News and SST-2.

Model	Robust	AG-News		SST-2	
		Acc.	Adv.	Acc.	Adv.
CLIP-ViT-L/14	✗	74.4	1.70	71.2	0.57
	✓	78.0	1.70	71.9	0.80
OpenCLIP-ViT-H/14	✗	71.1	1.60	61.6	1.83
	✓	72.3	1.00	58.4	2.98
OpenCLIP-ViT-g/14	✗	67.3	0.50	57.8	1.83
	✓	66.7	1.20	56.0	3.10