Zero-Shot Robustness of Vision Language Models Via Confidence-Aware Weighting

Nikoo Naghavian
School of ECE, College of Engineering
University of Tehran
nikoo.naghavian@ut.ac.ir

Mostafa Tavassolipour
School of ECE, College of Engineering
University of Tehran
tavassolipour@ut.ac.ir

Abstract

Vision-language models like CLIP demonstrate impressive zero-shot generalization but remain highly vulnerable to adversarial attacks. In this work, we propose Confidence-Aware Weighting (CAW) to enhance zero-shot robustness in vision-language models. CAW consists of two components: (1) a Confidence-Aware loss that prioritizes uncertain adversarial examples by scaling the KL divergence between clean and adversarial predictions, and (2) a feature alignment regularization that preserves semantic consistency by minimizing the distance between frozen and fine-tuned image encoder features on adversarial inputs. These components work jointly to improve both clean and robust accuracy without sacrificing generalization. Extensive experiments on TinyImageNet and 14 additional datasets show that CAW outperforms recent methods such as PMG-AFT and TGA-ZSR under strong attacks like AutoAttack, while using less memory.

1 Introduction

Traditional deep learning approaches rely on pre-training followed by fine-tuning with labeled data for each downstream task. The emergence of GPT-3 [1] in the natural language processing field has popularized models with zero-shot capability, where models trained on diverse internet-scale data can be applied to a wide range of tasks and unseen domains. In the multimodal setting, CLIP [2] employs a contrastive loss [3, 4] to align matching image–text pairs in a shared embedding space while separating mismatched pairs. This enables the model to acquire broad vision–language knowledge and achieve strong performance across various tasks, including image classification [5, 6], semantic segmentation [7], object detection [8, 9], image–text retrieval [10], and visual question answering [11]. Although CLIP demonstrates strong generalization ability, it remains vulnerable to small, imperceptible perturbations that leave the image visually unchanged to humans but cause significant shifts in predictions [12]. Adversarial training [13] is among the most effective approaches for improving robustness against strong attacks, typically training from scratch with both adversarial and clean examples. However, when applied to large-scale models like CLIP, adversarial training must be adapted to prevent overfitting and the forgetting of pre-trained knowledge, while still enhancing robustness [14, 15, 16].

The TeCoA [16] method was the first to study the zero-shot robustness of large-scale vision-language models. It showed the importance of using text supervision with a contrastive adversarial loss while applying different adaptation approaches [17]. Later, the PMG-AFT [18] method added new terms to the previous loss function to enhance robustness while causing a smaller decrease in performance on clean data. More recently, TGA-ZSR [19] introduced a method that improves both robustness and clean accuracy, along with the interpretability of attacks. This approach used text supervision with semantic information instead of relying on the model's output probabilities. Despite their

Accepted to the NeurIPS 2025 Workshop on Reliable ML from Unreliable Data.

effectiveness, these methods either need high memory usage, or still struggle to maintain robust accuracy under strong attacks.

We propose a novel adversarial fine-tuning loss named Confidence-Aware Weighting (CAW) that improves the robustness of a pre-trained CLIP model while preserving clean accuracy and reducing memory usage. This method introduces two key components designed to improve robustness and maintain generalization. The first is a Confidence-Aware term, which weights the KL divergence between clean and adversarial prediction distributions of the fine-tuned and frozen pre-trained CLIP models, ensuring that training focuses more on hard adversarial examples. The second is a regularization term, which matches adversarial image features from the fine-tuned image encoder with those from the frozen pre-trained encoder, helping retain semantic knowledge from the pre-trained model and reducing overfitting. Experiments on TinyImageNet and 14 zero-shot datasets (see Appendix B for details) demonstrate state-of-the-art performance under AutoAttack, surpassing both PMG-AFT and TGA-ZSR in robust accuracy. Under PGD-100 and CW, the proposed method outperforms PMG-AFT in both robust and clean accuracy, while maintaining lower memory usage than both baselines.

The key contributions of this work are:

- Propose CAW to improve zero-shot robustness by emphasizing challenging samples.
- Achieves higher robust accuracy than PMG-AFT and TGA-ZSR under AutoAttack.
- Improves clean and robust accuracy over PMG-AFT under PGD-100 and CW.
- Requires less memory than PMG-AFT and TGA-ZSR.

2 Methodology

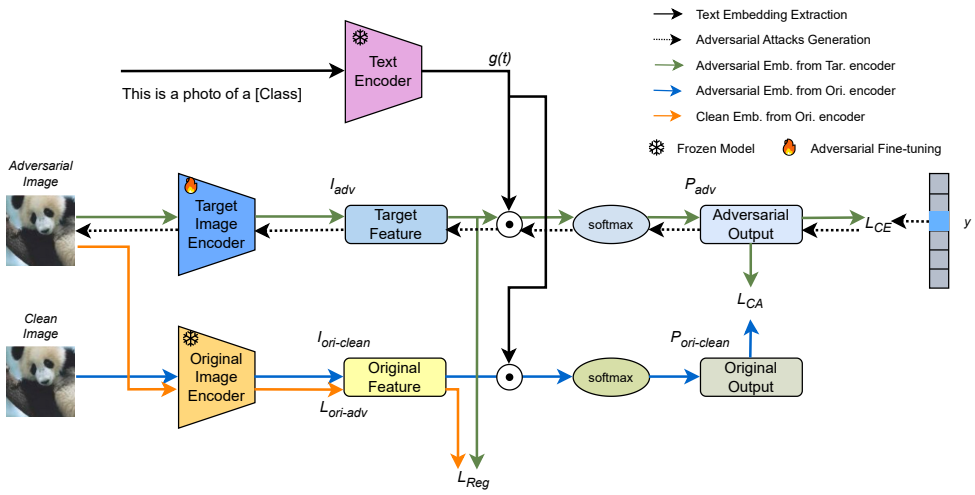

Figure 1: Overview of Confidence-Aware Weighting (CAW) method. $\odot$ means matrix inner product.

2.1 Preliminaries and Problem Setup

In this work, we employ CLIP [2] to enhance zero-shot robustness in classification tasks. CLIP has two encoders that learn a joint visual–text feature space. At inference, the predicted label is the one that text embedding has the highest cosine similarity with the image embedding. Following prior works [19, 18], we fine-tune the model using the cross-entropy loss:

$$L_{CE}(x,t,y) = -\mathbb{E}_{i,j}\left[y_{ij}\log\frac{\exp(\cos(f(x)_i, g(t)_j)/\tau)}{\sum_k \exp(\cos(f(x)_i, g(t)_k)/\tau)}\right] \tag{1}$$

where $f(x)$ and $g(t)$ denote the image and text embeddings, τ is the temperature parameter, and $\cos$ is the cosine similarity. The label y_{ij} is set to 1 for positive image-text pairs and 0 otherwise.

2.1.1 Adversarial Attacks

Deep learning models are typically trained and evaluated on clean data; however, small, imperceptible perturbations can cause significant prediction errors. Such perturbations can be generated using attack methods including PGD [20], AutoAttack [21], and CW [22]. PGD is an iterative attack that applies noise over multiple steps, producing stronger adversarial examples x_a than single-step methods such as FGSM [12]. It seeks perturbations that maximize the loss while keeping the perturbed input within a specified neighborhood of the clean example:

$$x_{a+1} = \Pi_{x+\mathcal{S}} \left(x_a + \varepsilon \cdot \operatorname{sign} \left(\nabla_{x_a} L(x_a, t, y) \right) \right) \tag{2}$$

where L denotes the loss function, x is the clean input, ε is the perturbation bound under the p-norm, and $\nabla_x L$ is the gradient direction that increases the loss. The set $\mathcal{S}$ represents the allowed changes that the adversary can make to the input.

2.1.2 Adversarial Training

By optimizing over adversarially perturbed inputs, adversarial fine-tuning enables models to learn more robust features through a min-max objective. In some cases, including our method, the objective function used to craft adversarial examples differs from the one used to optimize the model parameters. Specifically, in the inner loop (Equation 3), adversarial examples x_a are generated by maximizing the loss L (i.e., $\mathcal{L}_{\text{CE}}$, as defined in Equation 1 of our method), which is optimized using the PGD update rule (Equation 2). In the outer loop (Equation 4), the model parameters θ are updated by minimizing a separate loss function $\mathcal{J}$ (i.e., $\mathcal{L}_{\text{total}}$, as defined in Equation 10 of our method).

$$x_a = \arg\max_x \; L(x, t, y) \tag{3}$$

$$\theta^\star = \arg\min_\theta \; \mathcal{J}_\theta(x_a, t, y) \tag{4}$$

2.2 Method

Building on previous studies [16, 18, 19], we aim to preserve the generalizable and robust features learned by the pre-trained CLIP model during fine-tuning with a new loss function. As illustrated in Figure 1, we use both the original and target image encoders to retain prior knowledge while improving robustness. Although the TeCoA method [16] introduces a contrastive loss using adversarial examples with text supervision, it remains insufficient for jointly improving clean and robust accuracy. To address this limitation, we propose two additional loss terms that enhance robustness while maintaining generalization to unseen tasks.

Confidence-Aware Term We propose Confidence-Aware loss that focuses on challenging samples by emphasizing hard adversarial examples, i.e., those where the model is less confident in the correct class, while down-weighting easier ones. In contrast to prior methods that treat all samples equally in the loss function, our approach explicitly targets the inherent weaknesses in adversarial training by assigning more weight to samples that are more easily fooled by adversaries. This idea is inspired by the ARoW method [23], which prioritizes vulnerable samples to enhance adversarial robustness. However, our formulation significantly differs in both design and scope, as it is tailored to the unique challenges of vision-language models and zero-shot generalization. Specifically, we define a KL-based alignment between the frozen CLIP model's predictions on clean images, P^{clean}, and the fine-tuned model's predictions on adversarial images, P^{adv}. This alignment allows the model to retain semantic knowledge from pre-training while learning to handle difficult adversarial examples. Unlike ARoW, which uses the reverse KL divergence ($\text{KL}(P^{\text{clean}} \| P^{\text{adv}})$), we place the adversarial distribution as the first argument, i.e., $\text{KL}(P^{\text{adv}} \| P^{\text{clean}})$, which showed better results in our experiments. The distributions P^{adv} and P^{clean} are defined as:

$$P^{\text{adv}} = \text{softmax}(f(x_{\text{adv}})_{\text{tar}} \cdot g(t)^\top), \tag{5}$$

$$P^{\text{clean}} = \text{softmax}(f(x_{\text{clean}})_{\text{ori}} \cdot g(t)^\top), \tag{6}$$

where $f(\cdot)$ and $g(\cdot)$ denote the image and text encoder embeddings, and the dot operator represents the matrix inner product between these embeddings. The subscripts `tar` and `ori` refer to features from the fine-tuned and frozen image encoders. The element P^{adv}_{i,y_i} denotes the predicted probability for the true label y_i under the adversarial input x_i^{adv}, as defined in Equation 7:

$$P^{\text{adv}}_{i,y_i} = \left[\text{softmax} \left(f(x_i^{\text{adv}}) \cdot g(t)^\top \right) \right]_{y_i}. \tag{7}$$

Table 1: Zero-shot robust accuracy under AutoAttack with $\epsilon = 1/255$ on 15 datasets. We highlight the optimal accuracy in bold and underline the second-best result.

Methods	Tiny-ImageNet	CIFAR-10	CIFAR-100	STL-10	SUN397	Food101	OxfordPets	Flowers102	DTD	EuroSAT	FGVCAircraft	Caltech-101	Caltech-256	StanfordCars	PCAM	Average
CLIP	0.02	0.01	0.08	0.03	0.04	0.01	0.00	0.03	0.16	0.12	0.06	0.43	0.10	0.11	0.22	0.09
FT-Clean	0.08	0.03	0.01	0.91	0.09	0.04	0.06	0.03	0.48	0.02	0.03	1.38	0.66	0.03	0.03	0.26
FT-Adv	50.48	37.55	20.39	69.14	16.25	11.23	33.91	18.54	**19.95**	11.59	1.65	49.90	39.24	7.57	**48.84**	29.08
TeCoA	35.03	28.18	16.09	66.08	17.41	13.05	34.81	20.80	15.37	11.40	1.32	54.54	40.15	7.15	47.12	27.23
FARE	28.59	23.37	13.58	60.70	9.72	13.88	27.72	15.48	9.15	0.25	0.87	47.45	36.68	6.77	10.23	20.30
PMG-AFT	44.26	44.12	23.66	73.90	19.63	17.25	39.25	20.87	13.72	**11.99**	1.68	60.57	44.25	9.59	48.53	31.55
TGA-ZSR	49.45	40.53	22.38	72.06	**20.36**	15.58	40.31	21.43	17.13	11.19	**2.64**	57.16	45.68	10.47	48.03	31.63
CAW	**50.52**	**47.35**	**26.35**	**74.27**	19.64	**20.50**	**41.89**	**21.61**	16.80	11.11	2.52	**62.79**	**47.27**	**12.23**	47.81	**33.51**

To incorporate this into training, we minimize the KL divergence between P^{adv} and P^{clean}, scaled by $1 - P^{\text{adv}}_{i,y_i}$ to give greater importance to uncertain adversarial examples. This results in the Confidence-Aware loss:

$$L_{\text{CA}} = \frac{1}{N} \sum_{i=1}^{N} \left[\text{KL} \left(P_i^{\text{adv}} \parallel P_i^{\text{clean}} \right) \left(1 - P_{i,y_i}^{\text{adv}} \right) \right]. \tag{8}$$

Regularization Term We introduce a regularization loss that encourages consistency between the image encoder features of the frozen model, $f(\cdot)_{\text{ori}}$, and the fine-tuned model, $f(\cdot)_{\text{tar}}$, for adversarial inputs. This loss is computed before the text alignment stage, where the image features contain rich semantic information about the visual input. By aligning these features using the ℓ_2 distance metric, the model retains the pre-trained CLIP knowledge and reduces the risk of overfitting during adversarial fine-tuning. The regularization loss is defined as:

$$L_{\text{Reg}} = \frac{1}{N} \sum_{i=0}^{N} \| f(x_{\text{adv}})_{\text{tar}} - f(x_{\text{adv}})_{\text{ori}} \|_2 . \tag{9}$$

The overall loss function is formulated as follows:

$$L_{\text{total}} = L_{\text{CE}} + \alpha \cdot L_{\text{CA}} + \beta \cdot L_{\text{Reg}}. \tag{10}$$

3 Experiments

AutoAttack As shown in Table 1, our method outperforms all compared approaches under AutoAttack. On average, it achieves a 2% improvement in robust accuracy, demonstrating that the proposed training strategy learns transferable and more robust features resistant to this stronger attack. The model is trained with PGD-2 using a perturbation bound of $\epsilon = 1/255$ and evaluated on AutoAttack with the same perturbation bound. See Appendix A for related work, Appendix B for implementation details and datasets, Appendix C for additional experiments and ablation studies, and Appendix D for limitations and broader impact.

4 Conclusion

In this work, we demonstrate that emphasizing vulnerable samples during training improves the zero-shot robustness of CLIP. To this end, we introduce a CAW method that encourages the model to focus on hard adversarial examples, enabling the learning of more robust and transferable features. Experimental results show that our method outperforms prior approaches in both clean and robust accuracy across diverse domains under strong attacks, while requiring less memory, which is important for large-scale models. For future work, we aim to design a loss function that combines the idea of weighting challenging samples with attention mechanisms, which are essential components of large-scale models, to achieve better robustness against various attacks. Additionally, improving model interpretability by analyzing the features that contribute to robustness on difficult examples may provide deeper insights into building more resilient vision-language models.

References

[1] Ben Mann, Nick Ryder, Melanie Subbiah, J Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, S Agarwal, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 1(3):3, 2020.

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

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

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

[5] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

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

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848, 2017.

[8] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[9] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[10] Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. Image-text retrieval: A survey on recent research and development. *arXiv preprint arXiv:2203.14713*, 2022.

[11] Maria Parelli, Alexandros Delitzas, Nikolas Hars, Georgios Vlassis, Sotirios Anagnostidis, Gregor Bachmann, and Thomas Hofmann. Clip-guided vision-language pre-training for question answering in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5607–5612, 2023.

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

[13] 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*, 32, 2019.

[14] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems*, 36:54111–54138, 2023.

[15] 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*, pages 7472–7482. PMLR, 2019.

[16] Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. *arXiv preprint arXiv:2212.07016*, 2022.

[17] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022.

[18] Sibo Wang, Jie Zhang, Zheng Yuan, and Shiguang Shan. Pre-trained model guided fine-tuning for zero-shot adversarial robustness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24502–24511, 2024.

[19] Lu Yu, Haiyang Zhang, and Changsheng Xu. Text-guided attention is all you need for zero-shot robustness in vision-language models. *Advances in Neural Information Processing Systems*, 37: 96424–96448, 2024.

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

[21] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020.

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

[23] Dongyoon Yang, Insung Kong, and Yongdai Kim. Improving adversarial robustness by putting more regularizations on less robust samples. In *International Conference on Machine Learning*, pages 39331–39348. PMLR, 2023.

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

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

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

[27] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2020.

[28] Qi Liu, Tao Liu, Zihao Liu, Yanzhi Wang, Yier Jin, and Wujie Wen. Security analysis and enhancement of model compressed deep learning systems under adversarial attacks. In *2018 23rd Asia and South Pacific Design Automation Conference (ASP-DAC)*, pages 721–726. IEEE, 2018.

[29] Guneet S Dhillon, Kamyar Azizzadenesheli, Zachary C Lipton, Jeremy Bernstein, Jean Kossaifi, Aran Khanna, and Anima Anandkumar. Stochastic activation pruning for robust adversarial defense. *arXiv preprint arXiv:1803.01442*, 2018.

[30] Shixiang Gu and Luca Rigazio. Towards deep neural network architectures robust to adversarial examples. *arXiv preprint arXiv:1412.5068*, 2014.

[31] Andrew Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

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

[33] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International conference on learning representations*, 2019.

[34] Rahul Rade and Seyed-Mohsen Moosavi-Dezfooli. Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In *International Conference on Learning Representations*, 2022.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[36] 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)*, pages 4171–4186, 2019.

[37] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[38] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

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

[40] Xiao Li, Wei Zhang, Yining Liu, Zhanhao Hu, Bo Zhang, and Xiaolin Hu. Language-driven anchors for zero-shot adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24686–24695, 2024.

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

[42] Junhao Dong, Piotr Koniusz, Yifei Zhang, Hao Zhu, Weiming Liu, Xinghua Qu, and Yew-Soon Ong. Improving zero-shot adversarial robustness in vision-language models by closed-form alignment of adversarial path simplices. In *Forty-second International Conference on Machine Learning*.

[43] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[44] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[45] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

[46] Li Fei-Fei, Robert Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.

[47] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007.

[48] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505, 2012. doi: 10.1109/CVPR.2012.6248092.

[49] 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*, pages 722–729. IEEE, 2008.

[50] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[51] 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*, pages 554–561, 2013.

[52] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.

[53] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014.

[54] 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):2217–2226, 2019.

[55] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.

[56] Babak Ehteshami Bejnordi, Mitko Veta, Paul J. van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A. W. M. van der Laak, and the CAMELYON16 Consortium. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, 2017. doi: 10.1001/jama.2017.14585. URL `https://doi.org/10.1001/jama.2017.14585`.

Appendix

A Related Work

Adversarial robustness Deep neural networks have achieved remarkable performance on complex tasks, often producing highly confident predictions. However, small, imperceptible perturbations to the input can easily mislead them, resulting in incorrect outputs [20, 24, 25, 12, 26]. To address this vulnerability, various techniques have been proposed, including distillation [27], model compression [28], activation pruning [29], gradient regularization [30, 31] and adversarial training [20]. Adversarial training remains the most effective approach, augmenting adversarial examples alongside clean data during training to improve robustness while maintaining generalization [32]. Methods such as TRADES [15] balance clean accuracy and robustness by combining standard classification loss with a robustness regularization term. MART [33] highlights the importance of misclassified examples for improving robustness, while ARoW [23] focuses on the most vulnerable samples to enhance both generalization and robustness. HAT [34] mitigates over-robustness by introducing helper examples, arguing that pushing decision boundaries too far can harm clean accuracy.

Zero-shot Adversarial Robustness for VLMs The introduction of the attention mechanism [35], combined with advances in GPUs and access to large-scale unlabeled internet data, enabled the development of language models like BERT [36], GPT-2 [37], and GPT-3 [1], marking a new era in deep learning. GPT-3's emergence brought zero-shot capabilities, allowing knowledge transfer to unseen domains and tasks. Following this trend, vision-language models (VLMs) such as BLIP [38], CLIP [2] and ALIGN [39] incorporate textual information with images to improve performance across diverse tasks rather than a single downstream application. Despite their generalization ability, VLMs remain vulnerable to imperceptible perturbations in the input, which can cause incorrect predictions[14]. Recent research has explored enhancing VLM robustness. TeCoA [16] introduced the use of text knowledge for model alignment with adversarial examples through contrastive loss. PMG-AFT [18] and TGA-ZSR [19] extended this approach by adding terms such as KL divergence or semantic alignment with text embeddings to improve both clean and robust accuracy. Another method [40] extracts normalized semantic feature embeddings (anchors) for each class label from a CLIP text encoder and uses them to guide the image encoder during adversarial training, enabling robustness transfer to unseen categories. FARE approach [41] aligns adversarial example features directly with the embeddings of a pre-trained CLIP model without requiring labels. Another related work [42] leverages not only the final adversarial examples from the PGD process but also intermediate samples along the adversarial trajectory for training. Our work focuses on challenging adversarial examples to guide the model in learning more robust and generalizable features.

B Datasets and Implementation

Datasets To evaluate both clean and adversarial performance, we conduct extensive experiments on a diverse collection of image classification datasets. Our primary model, a pre-trained CLIP, is fine-tuned on the TinyImageNet [43] dataset. Evaluation is then performed not only on TinyImageNet but also on 14 additional datasets spanning five distinct domains. These include general object recognition benchmarks such as CIFAR-10 [44], CIFAR-100 [44], STL-10 [45], Caltech-101 [46], and Caltech-256 [47]; fine-grained classification datasets like OxfordPets [48], Flowers102 [49], FGVCAircraft [50], and StanfordCars [51]; scene recognition via SUN397 [52]; domain-specific datasets including Food101 [53], EuroSAT [54], and DTD [55]; and one medical imaging dataset, PCAM [56].

Implementation Details For implementation, we use the ViT-B/32 architecture as the backbone for the CLIP model and fine-tune it on adversarial examples generated from the TinyImageNet dataset. Adversarial examples for both training and evaluation are produced using PGD attacks under the ℓ_∞ norm. Training updates all image encoder parameters using SGD with a learning rate of 1×10^{-4}, momentum of 0.9, weight decay of 0, and a batch size of 128. All experiments use PGD with 2 iterations and a perturbation bound of $1/255$. For evaluation, we employ PGD-100, AutoAttack, and CW, each with a step size equal to the perturbation bound. We set the hyperparameters $\alpha = 6$ and $\beta = 3$ to balance clean and robust accuracy. To ensure fair comparison, we adopt settings consistent with prior studies, which also used an RTX 3090 GPU.

Table 2: Zero-shot robust accuracy under PGD-100 with $\epsilon = 1/255$ on 15 datasets. All methods are fine-tuned on TinyImageNet using PGD-2.

Methods	Tiny-ImageNet	CIFAR-10	CIFAR-100	STL-10	SUN397	Food101	OxfordPets	Flowers102	DTD	EuroSAT	FGVCAircraft	Caltech-101	Caltech-256	StanfordCars	PCAM	Average
CLIP	0.88	2.42	0.26	26.11	1.00	6.60	3.84	1.19	2.02	0.05	0.00	19.88	12.60	0.20	0.11	5.14
FT-Clean	13.55	19.92	4.94	40.00	0.82	2.64	2.40	0.68	2.66	0.05	0.03	14.95	9.69	0.09	1.32	7.58
FT-Adv.	51.59	38.58	21.28	69.55	17.60	12.55	34.97	19.92	15.90	11.95	1.83	50.73	48.48	8.42	**48.88**	30.15
TeCoA	37.57	30.30	17.53	69.17	19.70	14.76	36.44	22.46	17.45	12.14	1.62	55.86	41.89	8.79	47.39	28.87
FARE	23.88	21.25	10.72	59.59	8.30	10.97	24.56	15.48	10.96	0.14	0.84	45.96	34.35	4.38	10.17	18.77
PMG-AFT	47.11	46.01	25.83	73.92	**22.21**	19.58	41.62	23.45	15.05	**12.54**	1.98	62.42	45.99	11.72	48.64	33.20
CAW	**52.16**	**48.21**	**27.99**	**74.83**	21.33	**22.72**	**43.41**	**24.06**	**18.24**	11.93	**3.51**	**63.99**	**48.68**	**14.68**	47.92	**34.91**

Table 3: Zero-shot clean accuracy under PGD-100 with $\epsilon = 1/255$ on 15 datasets. All methods are fine-tuned on TinyImageNet using PGD-2.

Methods	Tiny-ImageNet	CIFAR-10	CIFAR-100	STL-10	SUN397	Food101	OxfordPets	Flowers102	DTD	EuroSAT	FGVCAircraft	Caltech-101	Caltech-256	StanfordCars	PCAM	Average
CLIP	57.26	**88.06**	60.45	**97.04**	**57.26**	**83.89**	**87.41**	**65.47**	**40.69**	**42.59**	**20.25**	**85.34**	**81.73**	**52.02**	**52.09**	**64.77**
FT-Clean	**79.04**	84.55	54.25	93.78	46.80	80.98	46.33	30.32	24.39	9.30	9.30	78.69	70.81	31.15	47.89	52.51
FT-Adv	73.83	68.96	39.69	86.89	33.37	27.74	60.10	33.45	13.26	16.49	4.86	67.41	57.72	18.11	49.91	43.45
TeCoA	63.97	66.14	36.74	87.24	40.54	35.11	66.15	33.25	13.75	17.13	6.75	64.63	56.20	25.65	49.01	44.15
FARE	77.54	87.58	**62.80**	94.33	49.91	70.02	81.47	57.10	36.33	22.69	14.19	84.04	77.50	44.35	46.07	60.39
PMG-AFT	67.11	74.62	44.68	88.85	37.42	37.47	66.34	35.66	21.17	17.76	4.71	76.70	61.96	25.21	49.60	47.28
CAW	75.64	82.96	55.49	91.36	41.96	50.87	71.02	42.15	28.56	23.42	9.42	80.66	67.94	34.88	49.98	53.75

C Ablation studies

To evaluate our method, we compare against the reported results of CLIP, FT-Clean, FT-Adv, TeCoA, FARE [41], PMG-AFT, and TGA-ZSR, as presented in the TGA-ZSR paper [19]. FT-Clean and FT-Adv are fine-tuned using clean and adversarial examples, both with contrastive loss.

PGD and CW Attack As shown in Table 2, our method outperforms PMG-AFT in robust accuracy on most datasets, achieving a higher average performance. Table 3 further demonstrates that our method surpasses PMG-AFT in clean accuracy across all datasets. Based on these results, our method performs well on both clean and adversarial samples, showing competitive performance compared to other approaches. Table 4 indicates that our approach achieves better results than PMG-AFT under the CW attack. We compare only with CLIP and PMG-AFT because these are the only methods reported in the paper [19]. The model is trained with PGD-2 using a perturbation bound of $\epsilon = 1/255$ and evaluated on PGD-100 and CW with the same bound.

Effect of Attack Strength Table 5 presents the average robust accuracy under PGD-100 with perturbation bounds of $\epsilon = 1/255$, $2/255$, and $4/255$ across 15 datasets. Our method outperforms PMG-AFT on average and surpasses other baseline methods across various attack strengths.

Analyzing the Effect of Each Loss Component As shown in Table 6, the L_{CE} row reports the average clean and robust accuracy across all 15 datasets under PGD-100 with $\epsilon = 1/255$. The L_{CA} row presents the results after adding this component to the previous loss term. Finally, the L_{Reg} row reflects the performance using the full loss function. These results demonstrate that our method improves both robustness and clean accuracy on average, compared to the standard CLIP loss.

Analysis of Computational Cost and Memory Usage As shown in Table 7, our method uses less memory than both PMG-AFT and TGA-ZSR while achieving better accuracy under stronger attacks, as discussed in previous sections. It also maintains a training time comparable to the aforementioned approaches.

Table 4: Zero-shot robust accuracy under CW attack on 15 datasets. All methods are fine-tuned on TinyImageNet using PGD-2.

Methods	Tiny-ImageNet	CIFAR-10	CIFAR-100	STL-10	SUN397	Food101	OxfordPets	Flowers102	DTD	EuroSAT	FGVCAircraft	Caltech-101	Caltech-256	StanfordCars	PCAM	Average
CLIP	0.21	0.36	0.10	10.59	1.16	0.82	1.23	1.09	2.18	0.01	0.00	13.50	7.36	2.36	0.07	2.45
PMG-AFT	44.59	44.86	24.15	74.11	19.99	17.33	39.88	20.95	13.51	**12.09**	1.47	60.99	44.46	10.57	**48.59**	32.36
CAW	**51.7**	**47.68**	**26.80**	**74.62**	**20.46**	**21.52**	**43.79**	**22.29**	**16.22**	11.60	**3.51**	**63.48**	**47.91**	**14.09**	47.71	**34.87**

Table 5: Zero-shot robust accuracy under PGD-100 with $\epsilon = 1/255, 2/255$ and $4/255$ on 15 datasets. All methods are fine-tuned on TinyImageNet using PGD-2.

Methods	Tiny-ImageNet	CIFAR-10	CIFAR-100	STL-10	SUN397	Food101	OxfordPets	Flowers102	DTD	EuroSAT	FGVCAircraft	Caltech-101	Caltech-256	StanfordCars	PCAM	Average
CLIP	0.64	2.15	0.12	20.35	0.52	5.94	2.97	0.72	0.71	0.03	0.00	14.28	9.18	0.11	0.04	3.65
FT-Clean	12.44	18.80	4.65	37.16	0.43	0.52	2.03	0.41	0.92	0.02	0.01	13.02	7.96	0.03	0.44	6.21
FT-Adv	29.33	18.10	11.06	45.13	8.58	5.65	16.45	10.15	9.72	9.82	0.83	33.43	24.14	3.80	38.06	17.07
TeCoA	18.17	12.78	8.12	39.87	8.53	6.12	11.04	10.07	10.07	9.88	0.63	34.94	23.92	3.45	33.20	15.41
FARE	12.41	9.09	4.23	33.72	2.98	4.75	9.67	5.52	4.26	0.25	0.28	23.97	16.95	1.48	3.43	8.54
PMG-AFT	25.30	21.71	13.29	47.69	**11.42**	9.49	**20.68**	**12.86**	9.45	**10.65**	0.90	**41.86**	28.92	3.72	**37.88**	19.27
CAW	**31.15**	**22.49**	**13.67**	**47.99**	9.87	**9.88**	20.16	12.14	**10.90**	7.05	**1.43**	41.33	**29.06**	**6.03**	29.88	**19.53**

D Discussion

Limitations Our method focuses solely on the CLIP model and has not been tested on other vision-language models under adversarial attacks. In addition, it only addresses adversarial perturbations in the image encoder, whereas the text encoder is also a crucial component of VLMs and should be considered to improve overall robustness.

Broder impact Large-scale models like VLMs have demonstrated strong zero-shot capabilities, performing well across diverse tasks and unseen domains. However, their performance under adversarial perturbations remains limited, which is an important and active area of research. As these models are increasingly deployed in real-world applications, ensuring their robustness and privacy against adversarial attacks becomes critical. Our method aims to improve the zero-shot robustness of CLIP under such attacks, contributing to the development of safer and more reliable vision-language systems.

Table 6: Average zero-shot robust and clean accuracy after adding each component, evaluated under PGD-100 with $\epsilon = 1/255$ on 15 datasets. All methods are fine-tuned on TinyImageNet using PGD-2.

	Robust	Clean	Average
CLIP	4.90	64.42	34.66
$\mathcal{L}_{CE}$	30.39	45.58	37.98
$+\mathcal{L}_{CA}$	33.64	51.50	42.57
$+\mathcal{L}_{Reg}$	34.92	53.65	44.28

Table 7: Memory Consumption and Training Time

Methods	Train memory usage	Train time (per epoch / batch)
CLIP	0Mb	0s / 0s
TeCoA	12,873Mb	512s / 0.65s
CAW	15,986Mb	842s / 1.08s
PMG-AFT	18,449Mb	828s / 1.06s
TGA-ZSR	21,227Mb	885s / 1.13s

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: The contributions of our method are clearly explained in both the introduction and abstract, and they are supported by experiments on 15 zero-shot datasets.

   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 the limitations of our method in Section D of the Appendix.

   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: Our work is entirely empirical and does not include theoretical assumptions or formal proofs.

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: Our method is described in full detail and step-by-step in Section 2.2, and comprehensive implementation details are provided in Appendix B, ensuring that the main experimental results can be reproduced.

   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: We will provide open access to the code.

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 provide full details of our experimental implementation, including training and evaluation settings, in Appendix B.

   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: We do not report error bars or statistical significance tests in this work.

   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: [Yes]

   Justification:We calculate the training memory usage and training time for our method and compare them with other baselines, as shown in Table 7.

   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: We have reviewed the NeurIPS Code of Ethics and confirm that our research aligns with its principles.

   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 discuss the broader impacts of our work in Section D of the Appendix.

    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: While our method builds on the publicly available CLIP model, we do not release any new pretrained model or dataset with high-risk misuse potential. Our work focuses on improving adversarial robustness using standard benchmarks, and no additional safeguards are necessary.

    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: The models, methodologies, datasets, and other elements used are appropriately aligned and referenced throughout the paper.

    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` 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: [NA]

    Justification: The paper does not release new assets.

    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: Our paper does not involve any crowdsourcing, user studies, or 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: Our paper does not involve any research with human subjects or participants, and therefore IRB approval is not applicable.

    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 rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: Our research does not involve LLMs in the core methodology; any usage was limited to writing or editing assistance.

    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.