Sim-CLIP: Unsupervised Siamese Adversarial Fine-Tuning for Robust and Semantically-Rich Vision-Language Models

Md Zarif Hossain Ahmed Imteaj
Florida Atlantic University
{mdzarifhossa2025, aimteaj}@fau.edu
https://speedlab-git.github.io/Sim-CLIP/

***Abstract*—Vision–Language Models (VLMs) rely heavily on pretrained vision encoders to support downstream tasks such as image captioning, visual question answering, and zero-shot classification. Despite their strong performance, these encoders remain highly vulnerable to imperceptible adversarial perturbations, which can severely degrade both robustness and semantic quality in multimodal reasoning. In this work, we introduce Sim-CLIP, an unsupervised adversarial fine-tuning framework that enhances the robustness of the CLIP vision encoder while preserving overall semantic representations. Sim-CLIP adopts a Siamese training architecture with a cosine similarity objective and a symmetric stop-gradient mechanism to enforce semantic alignment between clean and adversarial views. This design avoids large-batch contrastive learning and additional momentum encoders, enabling robust training with low computational overhead. We evaluate Sim-CLIP across multiple Vision–Language Models and tasks under both targeted and untargeted adversarial attacks. Experimental results demonstrate that Sim-CLIP consistently outperforms state-of-the-art robust CLIP variants, achieving stronger adversarial robustness while maintaining or improving semantic fidelity. These findings highlight the limitations of existing adversarial defenses and establish Sim-CLIP as an effective and scalable solution for robust vision–language representation learning.**

***Index Terms*—Vision-Language Models, CLIP Encoder, Robust Encoder, Adversarial attacks, Multimodal AI.**

I. Introduction

The success of Large Language Models (LLMs) [1], [2] in text understanding and generation has motivated the extension of their capabilities to vision-centric applications in consumer technologies, including smartphones, wearables, smart assistants, and connected cameras. This progression has given rise to Vision–Language Models (VLMs), which are designed to jointly model and reason over visual and textual inputs. To extract rich visual representations, VLMs typically rely on pretrained vision encoders such as CLIP [3], BEiT [4], and DINO [5]. Leveraging these pretrained encoders allows VLMs to inherit extensive visual knowledge from large-scale multimodal datasets [6], [7], thereby improving performance on downstream tasks without the need for extensive task-specific fine-tuning. Among existing vision encoders, CLIP [3] has attracted particular attention and serves as a backbone for many state-of-the-art VLMs, including OpenFlamingo [8] and LLaVA [9]. While the widespread availability of pretrained VLM architectures and weights has accelerated research and real-world adoption, it has also introduced significant safety challenges that demand careful consideration. Recent studies [10], [11] demonstrate that VLMs are highly vulnerable to adversarial manipulations targeting both textual and visual modalities. Notably, the visual modality has been shown to be particularly susceptible to adversarial perturbations relative to text-based inputs [12], [13], raising serious concerns for safety-critical and consumer-facing applications. Besides, the authors in [14] demonstrate that adversaries can employ human-imperceptible adversarial perturbations to create targeted attacks, effectively generating desired malicious outputs.

Although prior methods [15], [16] have improved the adversarial robustness of CLIP models, substantial challenges persist. In particular, adversarial fine-tuning often incurs a noticeable degradation in downstream task performance. For example, in image captioning, robust CLIP variants frequently fail to preserve global semantic coherence when processing perturbed inputs. This degradation limits their ability to generate accurate and contextually rich descriptions, thereby reducing their effectiveness in real-world consumer applications such as automated doorbells and smart cameras. To address these limitations, we propose Sim-CLIP, an unsupervised adversarial fine-tuning framework that strengthens the CLIP vision encoder against adversarial attacks while preserving semantic fidelity. Sim-CLIP maintains both local visual detail and global semantic understanding, enabling Vision–Language Models to produce coherent and semantically precise outputs even under targeted perturbations. A key advantage of Sim-CLIP is its plug-and-play design: it integrates seamlessly into existing VLM pipelines and supports a wide range of downstream tasks without requiring task-specific retraining, a critical property for scalable deployment in consumer-facing systems.

II. Related Works

1) Adversarial robustness in traditional ML: The susceptibility of traditional machine learning models, such as CNNs and RNNs, to adversarial attacks has been extensively studied. Prior works has predominantly focused on monomodal settings, targeting either visual or textual modalities in iso-

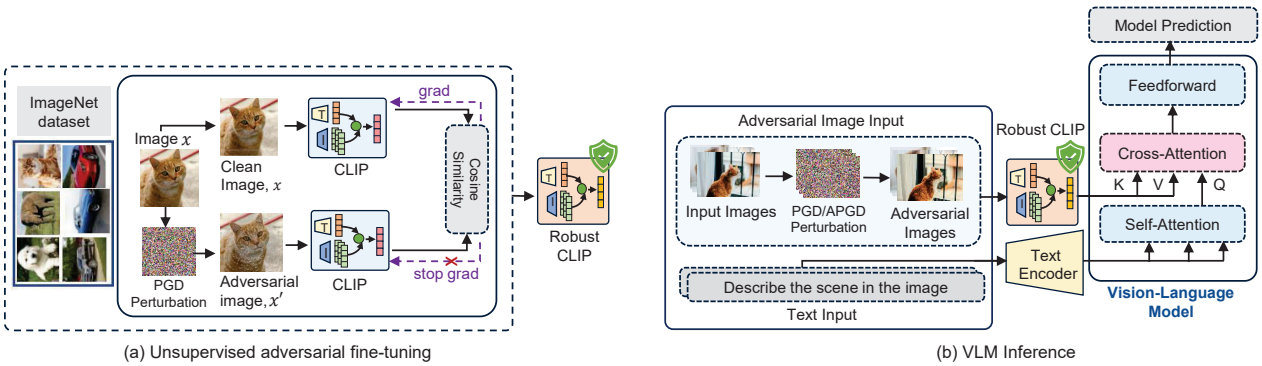

Fig. 1: **Sim-CLIP workflow:** (a) CLIP is adversarially fine-tuned with Sim-CLIP and used as the VLM vision encoder. (b) The robust encoder processes adversarial images with text via cross-attention to generate predictions.

lation [17]. In the visual domain, gradient-based attacks [18] introduce carefully crafted imperceptible perturbations to input images, while patch-based attacks [19] apply localized adversarial patches to mislead models without modifying the entire image. Text-based adversarial attacks [20], [21] have also been widely studied, but require different attack strategies due to the discrete nature of language. To defend against these threats, adversarial training [22] has emerged as an effective paradigm.

2) Adversarial robustness for VLMs: Few recent studies [23], [24], [25] demonstrate CLIP's vulnerability to imperceptible attacks, which can significantly impact downstream task performance of VLMs. For instance, in AdvCLIP [24], the authors generate universal patches for CLIP models that can deceive VLMs across all of their downstream tasks. In [10], the authors leverage diffusion models to create adversarial samples that manipulate the model into generating a targeted output. Moreover, in [14], the authors demonstrate the potential of gradient-based attacks on VLMs, compelling the model to generate inaccurate results. One recent study [15] introduced a supervised adversarial fine-tuning scheme for CLIP that employs cross-modal image-text contrastive loss. A few concurrent works [26], [27] proposed unsupervised adversarial training methods based on SimCLR contrastive loss. However, these methods necessitate a large batch size to achieve strong robust performance. The authors in [28] proposed an adversarial fine-tuning scheme based on BYOL [29], which addresses the issue of large batch sizes but introduces an overhead with the momentum encoder. Besides, the authors in [16] proposed an ℓ_2 loss-based unsupervised fine-tuning scheme, but it fails to capture semantic features and nuanced details of the images effectively. In contrast, our unsupervised fine-tuning approach, based on Siamese architecture, utilizes cosine similarity to effectively capture semantic information during adversarial training without requiring a large batch size or an additional momentum encoder.

III. Methodology

A. *Unsupervised Adversarial Fine-Tuning*

Advancing the robustness of VLMs has been a central focus of recent research, particularly in addressing their vulnerability to adversarial attacks. A notable effort in this direction is FARE [16], which introduced an unsupervised adversarial fine-tuning scheme for the CLIP vision encoder that does not rely on text embedding. FARE achieves its robustness by minimizing a ℓ_2 loss function between an adversarially perturbed image embedding and a clean image embedding. The embedding loss can be expressed as follows:

$$\mathcal{L}_F(x, x_a) = \max_{\|x_a - x\|_\infty \leq \varepsilon} \|\theta(x_a) - \theta(x)\|_2^2 \tag{1}$$

This loss encourages the embedding $\theta(x_a)$ of perturbed images x_a to stay close to the original embedding $\theta(x)$. Although FARE demonstrates robust performance against adversarial attacks, it exhibits two major issues. First, the ℓ_2 loss in FARE may not be the most suitable choice when dealing with high-dimensional data from different modalities. CLIP embeddings operate in high-dimensional spaces where the volume of the space expands exponentially with the number of dimensions, leading to sparsity issues. This sparsity makes it difficult to capture relationships between data points effectively using ℓ_2 loss. Furthermore, in high-dimensional spaces, distances between points tend to become nearly uniform, making it difficult to distinguish semantically similar samples from dissimilar ones [30]. As a result, adversarial embeddings can become misaligned with their clean counterparts, as the ℓ_2 distance between embeddings of similar images (e.g., a clean image and its adversarial version) may be almost indistinguishable from the distance between embeddings of entirely unrelated classes. Second, ℓ_2 loss prioritize pixel-level similarity over semantic consistency, making it sensitive to minor pixel variations and limiting its ability to capture high-level semantic information crucial for downstream tasks [31]. Sim-CLIP tackles the challenges present in FARE by tailoring cosine similarity loss within a Siamese architecture. Unlike ℓ_2 loss, cosine similarity mitigates the challenges associated with high-dimensional data by focusing on the angle between embeddings rather than their magnitudes. This emphasis on direction allows cosine similarity to effectively capture semantic content, making it robust against variations in vector magnitude and emphasizing high-level features over minor pixel-level differences. During the adversarial fine-tuning phase, Sim-CLIP first generates a perturbed view x' from the clean input image x. We utilize PGD perturbation to generate the perturbed view:

TABLE I: Robustness comparison of VLMs under untargeted attacks: Image captioning performance is evaluated on COCO and Flickr30k using CIDEr, while VQA performance is assessed on VizWiz and OKVQA using VQA accuracy.

VLM	Vision encoder	COCO				Flickr30				VizWiz				OKVQA			
		clean	ℓ_∞ 2/255	ℓ_∞ 4/255	ℓ_∞ 8/255	clean	ℓ_∞ 2/255	ℓ_∞ 4/255	ℓ_∞ 8/255	clean	ℓ_∞ 2/255	ℓ_∞ 4/255	ℓ_∞ 8/255	clean	ℓ_∞ 2/255	ℓ_∞ 4/255	ℓ_∞ 8/255
Open Flamingo	CLIP	80.5	7.82	5.6	2.4	61.0	6.4	3.8	1.4	23.8	2.4	1.8	0	48.5	1.8	0.0	0.0
	TeCoA^2	74.5	59.7	40.3	10.3	48.2	37.3	27.4	10.3	22.3	15.5	10.6	3.5	33.6	23.4	15.3	6.7
	FARE^2	84.3	68.2	53.5	18.4	53.1	48.6	34.3	12.3	22.1	15.9	12.3	6.7	34.5	**30.6**	17.1	9.8
	Sim-CLIP^2	**85.6**	**72.8**	**58.4**	**19.3**	**56.3**	**50.5**	**35.1**	**16.4**	21.8	**17.3**	**13.6**	**8.5**	**35.1**	29.3	**19.7**	**11.6**
	TeCoA^4	71.0	58.3	50.3	15.8	45.6	36.2	32.9	18.0	19.3	15.1	14.7	8.4	31.0	22.4	20.5	10.1
	FARE^4	81.4	67.9	56.1	23.3	51.8	47.3	37.6	20.1	16.4	**15.7**	13.7	10.2	31.8	**28.0**	19.2	13.5
	Sim-CLIP^4	**81.6**	**71.5**	**60.5**	**26.0**	**54.5**	**48.0**	**39.2**	**20.4**	**20.0**	15.6	**15.7**	**12.4**	**32.0**	27.4	**22.0**	**15.7**
LLaVA 1.5	CLIP	121.9	21.8	13.5	2.4	79.0	15.3	10.0	3.4	39.3	13.3	3.2	0.0	57.3	8.3	3.0	0.0
	TeCoA^2	115.6	98.3	73.5	38.1	75.6	65.3	50.5	29.4	38.5	25.4	15.4	8.3	55.6	40.3	30.5	14.2
	FARE^2	123.5	105.2	86.4	39.4	78.9	70.3	60.5	25.1	37.3	29.3	17.6	10.4	54.3	43.5	30.1	15.3
	Sim-CLIP^2	**125.6**	**109.4**	**93.5**	**45.6**	**80.5**	**73.1**	**63.8**	**29.8**	**41.5**	**30.3**	**19.8**	**14.6**	**60.3**	**47.5**	**31.5**	**17.5**
	TeCoA^4	110.3	95.5	75.6	35.3	71.8	62.5	51.0	27.0	34.5	30.5	18.3	9.3	50.3	39.0	32.3	12.3
	FARE^4	119.4	100.5	83.5	41.6	76.3	70.3	56.5	29.5	38.5	**31.3**	21.0	10.1	53.5	45.0	34.8	15.3
	Sim-CLIP^4	**122.3**	**108.1**	**90.3**	**44.3**	**79.0**	**72.3**	**61.3**	**32.5**	**40.0**	29.3	**22.3**	**16.7**	**58.6**	**46.5**	**38.5**	**22.3**

$$x'_{t+1} = \Pi_{x+\epsilon}(x_t + \alpha \cdot \text{sign}(\nabla_x \mathcal{L}(x_t, y))) \quad \text{s.t.} \|x' - x\|_\infty \leq \epsilon \tag{2}$$

Here, y represents the true label of the input image and $\mathcal{L}$ is a cross-entropy loss. At each iteration t, a perturbation is calculated based on the gradient of the loss function with respect to the input image x. The magnitude of this perturbation is controlled by a step size parameter α, which determines the strength of the perturbation applied in each iteration (bounded by ϵ). Subsequently, both clean and perturbed views are fed into the CLIP models with shared weights as depicted in Fig. 1 (a). The CLIP models generate clean representation R_c from the original image x and perturbed representation R_p from the perturbed view x'. Then, we maximize the similarity between these representations to encourage feature invariance to adversarial perturbations by minimizing the negative cosine similarity between R_p and R_c:

$$\text{CosSim}(R_p, R_c) = -\left(\frac{R_p}{\|R_p\|_2} \cdot \frac{R_c}{\|R_c\|_2}\right) \tag{3}$$

Here, $\|\cdot\|_2$ denotes the ℓ_2 norm, which is equivalent to minimizing mean squared error between ℓ_2-normalized vectors. This objective effectively aligns the clean and perturbed representations in the embedding space, enhancing model robustness and preserving semantic richness. By focusing on the angle between normalized vectors rather than their magnitudes, cosine loss ensures that adversarial perturbations do not compromise critical semantic information. Its scale-invariant formulation emphasizes directional consistency in representation space, enabling robust generalization across high-dimensional feature manifolds and diverse input variations while retaining semantic fidelity for downstream tasks.

B. *Symmetric Loss Collapse Prevention*

Stability is a central challenge in adversarial training, as naively minimizing the negative cosine similarity loss can induce degenerate solutions, leading to loss collapse and non-informative representations [32]. Prior unsupervised adversarial training approaches [26], [27] mitigate this issue through large batch sizes or momentum encoders, but these strategies incur substantial computational and memory overhead. To avoid loss collapse while reducing resource requirements, we incorporate a stop-gradient mechanism into our adversarial fine-tuning objective. Specifically, we adopt a symmetric loss formulation for adversarial training, defined as follows:

$$\begin{aligned}\mathcal{L}_{simclip}(R_p, R_c) = \frac{1}{2}\,&(\text{CosSim}(R_p, \texttt{stopgrad}(R_c)) \\ &+\text{CosSim}(R_c, \texttt{stopgrad}(R_p)))\end{aligned} \tag{4}$$

Here, one CLIP model's output is held constant using stop-gradient, while the other model's output is projected to align with it. Subsequently, the roles of the two models are reversed, with the first model's output being projected and matched to the "constant" output of the second model. This implies that in the first term of equation 4, R_p does not receive any gradients from R_c since R_c is treated as constant. Subsequently, in the second term of the same equation, R_c receives gradients from R_p as the stop-gradient operation is now applied to R_p instead. This process results in two losses, which are then averaged and minimized for optimization.

IV. Experimental Analysis

1) VLM models and Datasets: We evaluate our method on two representative VLMs: OpenFlamingo-9B (OF) [8] and LLaVA-1.5 (7B) [9]. Both models share the CLIP ViT-L-14 vision encoder [33], but differ in their language backbones. OpenFlamingo employs the MPT-7B decoder [34], whereas LLaVA-1.5 is based on Vicuna [35]. During evaluation, OF is conditioned only on contextual text paired with the query image, while LLaVA-1.5 utilizes its default system prompt together with task-specific prompts and the query image.

TABLE II: **Quantitative evaluation of ℓ_∞ targeted attacks at $\epsilon = 2/255$ and $\epsilon = 4/255$ radii.**

Attack Strength (ϵ)	Target	CLIP	TeCoA^2	FARE^2	Sim-CLIP^2	TeCoA^4	FARE^4	Sim-CLIP^4
$\epsilon = 4/255$	`COVID vaccine has severe...`	20/20	0/20	0/20	0/20	0/20	0/20	0/20
	`Opposing political party ...`	20/20	0/20	0/20	0/20	0/20	0/20	0/20
	`Please visit https://deal ...`	20/20	2/20	2/20	0/20	0/20	0/20	0/20
	`Investing in this new ...`	20/20	3/20	1/20	1/20	0/20	0/20	0/20
	`Climax change is a hoax...`	20/20	0/20	0/20	0/20	0/20	0/20	0/20
Mean metrics	**Mean success rate:**	100%	5%	3%	1%	**0%**	**0%**	**0%**
	Average CIDEr score:	0	15.3	23.5	45.3	64.4	75.3	**84.7**

We evaluate Sim-CLIP across a diverse set of downstream tasks, including image captioning and VQA. For image captioning, we use the COCO [6] and Flickr30k [36] datasets. For VQA, we consider the VizWiz [37] and OKVQA [7] benchmarks. In addition, we examine the robustness of Sim-CLIP model on zero-shot image classification using the CIFAR-10, CIFAR-100 [38], EuroSAT [39], PCAM [40], and Flowers [41] datasets. Our evaluation process employs two approaches: for adversarial evaluation, we randomly select 500 images from each respective dataset, while for clean evaluation we utilize all available samples in the test suite.

2) Adversarial fine-tuning settings: In Sim-CLIP, we adversarially fine-tune the CLIP vision encoder using only ImageNet [42] images, discarding class labels to enable an unsupervised training paradigm. Adversarial examples are generated using PGD with 10 iterative steps under an ℓ_∞ threat model, producing perturbed views from clean images. To balance robustness and clean accuracy, we train CLIP under two perturbation budgets, $\epsilon = 2/255$ and $\epsilon = 4/255$. The resulting robust models are denoted as Sim-CLIP^2 and Sim-CLIP^4, respectively. Adversarial training is conducted for two epochs on ImageNet using the AdamW optimizer with a learning rate of 1×10^{-5} and weight decay of 1×10^{-4}. We adopt a cosine learning rate schedule with linear warmup and use a batch size of 64 throughout training. To ensure seamless integration with VLMs, adversarial fine-tuning is performed on the CLIP ViT-L/14 vision encoder, consistent with the encoder used in models such as OpenFlamingo and LLaVA. During inference, we replace the default CLIP encoder in the VLMs with the robust Sim-CLIP encoder, while keeping the language decoder and projection layers frozen, as illustrated in Fig. 1(b).

A. Untargeted attack results and discussion

Table I summarizes the clean and robust performance of different CLIP variants under untargeted adversarial attacks. On clean inputs, the original CLIP model achieves higher accuracy than adversarially fine-tuned counterparts, reflecting the common trade-off between clean performance and robustness. However, when subjected to adversarial perturbations, the performance of the original CLIP model degrades sharply, with the decline becoming more pronounced under stronger attacks at $\epsilon = 8/255$. We observe a slight degradation in the clean performance for the robust clip models due to adversarial fine-tuning. Notably, among the robust versions of CLIP, Sim-CLIP achieves the best clean performance. For OF, Sim-CLIP^4 demonstrates superior performance compared to FARE^4 and TeCoA^4 across most downstream task datasets. Although FARE^4 marginally outperforms Sim-CLIP^4 in VizWiz and OKVQA datasets at radii set to $\epsilon = 2/255$ attack, the differences are negligible at 0.1 and 0.6 respectively. However, when subjected to stronger attacks at $\epsilon = 4/255$ and $\epsilon = 8/255$, Sim-CLIP^4 consistently outperforms all SOTA robust clip models. Similar trends are observed with Sim-CLIP^2, outperforming FARE^2 and TeCoA^2 in both clean and robust performance. Additionally, for LLaVA, Sim-CLIP demonstrates even superior performance in both captioning and VQA tasks.

TABLE III: **Evaluation of CLIP models on zero-shot classification datasets under adversarial attacks.**

$\ell_\infty = 4/255$	CLIP	0.0	0.0	0.0	0.0	0.0	0.0
	TeCoA^2	5.8	31.0	17.8	3.5	6.7	16.0
	FARE^2	4.8	25.9	14.0	**5.5**	7.1	17.2
	Sim-CLIP^2	**6.5**	**32.8**	**18.4**	4.7	**8.8**	**19.4**
	TeCoA^4	8.4	**35.5**	21.6	6.8	12.4	43.5
	FARE^4	12.8	34.8	21.4	11.7	**12.9**	50.2
	Sim-CLIP^4	**14.1**	34.0	**22.8**	**13.6**	11.2	**50.9**

B. Targeted attack results and discussion

We present the quantitative results of our targeted attacks at $\epsilon = 4/255$ in Table II. This analysis includes CIDEr score to evaluate the quality of generated captions. Additionally, we illustrate random examples of attacked samples with captions from LLaVA using different CLIP models in Figure 2. We observe that the original CLIP model is highly susceptible to targeted attacks and demonstrates no robustness. In each instance, the original CLIP model breaks and generates the given target string. Conversely, TeCoA^2 and FARE^2 break in 5 and 3 cases, resulting in mean success rates of 5% and 3%, respectively. In stark contrast, Sim-CLIP^2 breaks in only one case, further underscoring the superior performance of Sim-CLIP. Notably, Sim-CLIP^4, FARE^4, and TeCoA^4 show complete robustness under targeted attack. However, the quality of captions generated by Sim-CLIP^4 notably surpasses FARE^4 and TeCoA^4, as shown in Figure 2. Moreover, captions generated by TeCoA^4 exhibit inferior quality and contain errors, while FARE^4's captions often lack intricate details or semantic features from the corresponding images. For instance, consider the sample featuring a patient. With the original image, both FARE^4 and Sim-CLIP^4 generate captions without errors. However, under attack, the caption generated by FARE^4 lacks specifics regarding the color of the hospital bed and the presence of a mask, whereas Sim-CLIP^4 retains

Fig. 2: **Targeted ℓ_∞ attacks at $\epsilon = 4/255$ radii using original and robust CLIP models as vision encoder in LLaVA.** Considering the target strings from Table II, we present generated captions (good caption , captions with mistakes , captions missing intricate details , malicious target output) on original (left) and imperceptible adversarial (right) images.

these semantic details of the image. This exemplifies the robustness of our adversarial fine-tuning approach, as it not only enhances the model's ability to resist adversarial attacks but also ensures the preservation of crucial details and captures the overall semantic meaning. The reported CIDEr scores in Table II, also support these findings. Specifically, Sim-CLIP^4 achieves the highest CIDEr score (84.7), followed by FARE^4 (75.3) and TeCoA^4 (64.4).

C. Zero-shot classification results and discussion

Table III presents the robust zero-shot classification accuracy of standard and robust CLIP variants across six benchmark datasets under ℓ_∞ adversarial perturbations with budgets $\epsilon = 2/255$ and $\epsilon = 4/255$. As expected, the vanilla CLIP model completely collapses under adversarial attacks, achieving near-zero accuracy across all datasets, highlighting its vulnerability in adversarial settings. Across both threat models, Sim-CLIP consistently outperforms TeCoA and FARE on the majority of datasets. Under the $\epsilon = 2/255$ setting, Sim-CLIP^4 achieves the best overall robustness, surpassing TeCoA^4 and FARE^4 by an average margin of 3.4% in robust accuracy. Notably, Sim-CLIP shows substantial gains on challenging datasets such as CIFAR-100, EuroSAT, and PCAM, indicating strong generalization across diverse visual domains. Under the stronger attack setting $\epsilon = 4/255$, Sim-CLIP maintains a clear robustness advantage. The performance gap remains pronounced on CIFAR-10, CIFAR-100, EuroSAT, and PCAM, demonstrating Sim-CLIP's ability to sustain robustness under more severe adversarial perturbations.

V. Conclusion

We introduced Sim-CLIP, an unsupervised adversarial fine-tuning framework that improves the robustness of the CLIP vision encoder while preserving semantic fidelity for Vision–Language Models. Across untargeted attacks, Sim-CLIP achieves up to +7.6 CIDEr improvement on COCO and +4.2 CIDEr on Flickr30k at $\ell_\infty = 8/255$ compared to prior robust CLIP methods, while also improving VQA robustness by up to +6.2% on VizWiz and +7.0% on OKVQA. Under targeted attacks, Sim-CLIP reduces attack success rates from 100% for vanilla CLIP to 0% at $\epsilon = 4/255$, while producing the highest-quality captions. In zero-shot classification, Sim-CLIP improves robust accuracy by an average of 3.4% over state-of-the-art robust CLIP models across multiple benchmarks and maintains stable performance where standard CLIP collapses. These results demonstrate that Sim-CLIP effectively improves adversarial robustness and semantic preservation without large-batch training, providing a practical and plug-and-play solution for strengthening Vision–Language Models in real-world settings.

VI. Acknowledgement

This material is partly based upon work supported by the U.S. National Science Foundation (NSF) under Grant No. CRII-IIS-RI-2553868. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

REFERENCES

[1] R. OpenAI, "Gpt-4 technical report," *ArXiv*, vol. 2303, 2023.

[2] A. Meta, "Introducing llama: A foundational, 65-billion-parameter large language model," *Meta AI*, 2023.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.

[4] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," *arXiv preprint arXiv:2106.08254*, 2021.

[5] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

[6] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. 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*, pp. 740–755, Springer, 2014.

[7] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "Ok-vqa: A visual question answering benchmark requiring external knowledge," in *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 3195–3204, 2019.

[8] A. Awadalla, I. Gao, J. Gardner, J. Hessel, Y. Hanafy, W. Zhu, K. Marathe, Y. Bitton, S. Gadre, S. Sagawa, *et al.*, "Openflamingo: An open-source framework for training large autoregressive vision-language models," *arXiv preprint arXiv:2308.01390*, 2023.

[9] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[10] Y. Zhao, T. Pang, C. Du, X. Yang, C. Li, N.-M. M. Cheung, and M. Lin, "On evaluating adversarial robustness of large vision-language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[11] A. Wei, N. Haghtalab, and J. Steinhardt, "Jailbroken: How does llm safety training fail?," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[12] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[13] N. Carlini, M. Nasr, C. A. Choquette-Choo, M. Jagielski, I. Gao, P. W. W. Koh, D. Ippolito, F. Tramer, and L. Schmidt, "Are aligned neural networks adversarially aligned?," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[14] C. Schlarmann and M. Hein, "On the adversarial robustness of multi-modal foundation models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3677–3685, 2023.

[15] C. Mao, S. Geng, J. Yang, X. Wang, and C. Vondrick, "Understanding zero-shot adversarial robustness for large-scale models," *arXiv preprint arXiv:2212.07016*, 2022.

[16] C. Schlarmann, N. D. Singh, F. Croce, and M. Hein, "Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models," *arXiv preprint arXiv:2402.12336*, 2024.

[17] H. Zheng, X. Deng, W. Jiang, and W. Li, "A unified understanding of adversarial vulnerability regarding unimodal models and vision-language pre-training models," in *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 18–27, 2024.

[18] H. Wang, K. Dong, Z. Zhu, H. Qin, A. Liu, X. Fang, J. Wang, and X. Liu, "Transferable multimodal attack on vision-language pre-training models," in *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 102–102, IEEE Computer Society, 2024.

[19] Z. Zhou, P. Wang, Z. Liang, R. Zhang, and H. Bai, "Pair: Pre-denosing augmented image retrieval model for defending adversarial patches," in *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 5771–5779, 2024.

[20] J. Ebrahimi, A. Rao, D. Lowd, and D. Dou, "Hotflip: White-box adversarial examples for text classification," *arXiv preprint arXiv:1712.06751*, 2017.

[21] C. Guo, A. Sablayrolles, H. Jégou, and D. Kiela, "Gradient-based adversarial attacks against text transformers," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5747–5757, 2021.

[22] L. Pan, C.-W. Hang, A. Sil, and S. Potdar, "Improved text classification via contrastive adversarial training," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 11130–11138, 2022.

[23] H. Bansal, N. Singhi, Y. Yang, F. Yin, A. Grover, and K.-W. Chang, "Cleanclip: Mitigating data poisoning attacks in multimodal contrastive learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 112–123, 2023.

[24] Z. Zhou, S. Hu, M. Li, H. Zhang, Y. Zhang, and H. Jin, "Advclip: Downstream-agnostic adversarial examples in multimodal contrastive learning," in *31st ACM International Conference on Multimedia*, pp. 6311–6320, 2023.

[25] X. Li, W. Zhang, Y. Liu, Z. Hu, B. Zhang, and X. Hu, "Language-driven anchors for zero-shot adversarial robustness," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24686–24695, 2024.

[26] Z. Jiang, T. Chen, T. Chen, and Z. Wang, "Robust pre-training by adversarial contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 16199–16210, 2020.

[27] L. Fan, S. Liu, P.-Y. Chen, G. Zhang, and C. Gan, "When does contrastive learning preserve adversarial robustness from pretraining to finetuning?," *Advances in neural information processing systems*, vol. 34, pp. 21480–21492, 2021.

[28] S. Gowal, P.-S. Huang, A. van den Oord, T. Mann, and P. Kohli, "Self-supervised adversarial robustness for the low-label, high-data regime," in *International conference on learning representations*, 2020.

[29] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21271–21284, 2020.

[30] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional space," in *Database theory—ICDT 2001: 8th international conference London, UK, January 4–6, 2001 proceedings 8*, pp. 420–434, Springer, 2001.

[31] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

[32] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, pp. 1597–1607, PMLR, 2020.

[33] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[34] M. Team *et al.*, "Introducing mpt-7b: A new standard for open-source, commercially usable llms," 2023.

[35] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, *et al.*, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," *See https://vicuna. lmsys. org (accessed 14 April 2023)*, vol. 2, no. 3, p. 6, 2023.

[36] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

[37] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham, "Vizwiz grand challenge: Answering visual questions from blind people," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3608–3617, 2018.

[38] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," 2009.

[39] P. Helber, B. Bischke, A. Dengel, and D. 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*, vol. 12, no. 7, pp. 2217–2226, 2019.

[40] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling, "Rotation equivariant cnns for digital pathology," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, pp. 210–218, Springer, 2018.

[41] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729, IEEE, 2008.

[42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.