ADVERSARIAL ROBUSTNESS IN LARGE VISION-LANGUAGE MODELS: DETECTION, DEFENSE, AND CERTIFICATION

[1]Arul Selvan M
[1]Assistant Professor, Department of Computer Science and Engineering, K.L.N. College of Engineering, Sivaganga - 630612
Email : [1]arul2591@gmail.com

Abstract Adversarial robustness in large vision-language models (VLMs) has emerged as a critical research focus due to the growing deployment of these models in real-world applications, where their vulnerability to adversarial attacks can lead to severe consequences. These attacks exploit subtle, often imperceptible perturbations to input images or text, causing models to produce incorrect or misleading outputs, thereby undermining trust and reliability. This paper presents a comprehensive overview of adversarial robustness in VLMs, focusing on three fundamental aspects: detection, defense, and certification. First, detection methods aim to identify adversarial inputs before they influence the model's decision-making process. Techniques such as anomaly detection, input reconstruction, and model uncertainty estimation are discussed, highlighting their effectiveness and limitations in the vision-language domain. Next, defense strategies are explored, including adversarial training, input preprocessing, and robust architecture design, which seek to enhance the model's resilience against adversarial manipulations. We examine how these defenses can be tailored to the multi-modal nature of VLMs, addressing unique challenges such as the alignment of visual and textual modalities under attack. Additionally, we analyze emerging defense paradigms leveraging self-supervised learning and contrastive objectives that promote intrinsic robustness. Finally, certification approaches are reviewed, which provide theoretical guarantees on the robustness of VLMs within certain perturbation bounds, thereby offering provable assurance against adversarial examples. We discuss advances in randomized smoothing and verification techniques adapted for multi-modal inputs, emphasizing their role in establishing formal robustness benchmarks. Throughout the paper, we underscore the interplay between detection, defense, and certification, advocating for integrated frameworks that jointly address these facets to build more secure and reliable VLMs. We also identify key challenges and future directions, such as scalability to large-scale models, robustness to diverse and adaptive attack vectors, and the need for standardized evaluation protocols specific to vision-language tasks. By synthesizing recent developments and providing a holistic perspective, this work aims to guide researchers and practitioners in advancing adversarial robustness for large vision-language models, ultimately facilitating safer deployment in sensitive domains including autonomous systems, healthcare, and content moderation.

Keywords: adversarial robustness, vision-language models, adversarial detection, defense strategies, robustness certification, multi-modal learning

1. INTRODUCTION

In recent years, large vision-language models (VLMs) have revolutionized the field of artificial intelligence by bridging the gap between visual and textual data understanding. These models leverage large-scale pretraining on paired image-text datasets to learn rich multi-modal representations, enabling impressive performance in tasks such as image captioning, visual question answering, cross-modal retrieval, and zero-shot recognition. Examples of prominent VLMs include CLIP, ALIGN, and Flamingo, which have demonstrated remarkable ability to generalize across diverse visual and linguistic domains. However, despite their substantial success, the security and reliability of these models remain a critical concern. Like their unimodal counterparts in computer vision and natural language processing, VLMs are susceptible to adversarial attacks—carefully crafted inputs that cause the models to produce erroneous or misleading outputs without perceptible changes to human observers.

The vulnerability of machine learning models to adversarial perturbations poses a significant threat when such systems are deployed in real-world applications, especially in high-stakes or safety-critical domains such as autonomous driving, medical diagnosis, and content moderation. The multi-modal nature of VLMs introduces unique challenges and attack surfaces, as adversaries can manipulate either or both modalities (images and texts) to deceive the model. For example, slight modifications to an image or its associated textual context could lead to incorrect classification or retrieval results, undermining the model's reliability and user trust. This has galvanized an active research area focused on adversarial robustness, aiming to develop methods to detect adversarial inputs, defend models against attacks, and certify the robustness guarantees that quantify the limits of model susceptibility.

Adversarial robustness research in the vision and language domains has evolved along three main lines: detection, defense, and certification. Detection methods focus on identifying whether an input is adversarial before the model processes it or before decisions are made. This is particularly crucial in scenarios where outright prevention of attacks may be infeasible, but timely detection allows for rejection, alerting, or fallback mechanisms. Typical approaches include anomaly detection based on input distribution shifts, uncertainty quantification via Bayesian or ensemble methods, and input reconstruction techniques such as denoising or generative modeling. While such techniques have been extensively studied in unimodal settings, adapting them to vision-language models requires addressing the complexity of multi-modal feature interactions and the potential for cross-modal inconsistency induced by adversarial perturbations.

Defense strategies form the second pillar of adversarial robustness and seek to harden models against attacks by either modifying the training process, the model architecture, or the input data pipeline. Adversarial training, where models are exposed to adversarial examples during training, has proven effective in image classification tasks and has been extended to VLMs with promising results. However, the joint visual and textual modalities complicate adversarial training, as attacks may target one or both modalities, necessitating multi-modal adversarial examples and corresponding defense mechanisms. Other defense methods involve input preprocessing steps such as random resizing, compression, or feature smoothing that mitigate adversarial effects. Architectural innovations that promote robust representation learning, such as attention mechanisms that emphasize salient multi-modal cues, also contribute to enhanced defense. Furthermore, recent advances in self-supervised learning and contrastive learning objectives offer avenues to encourage intrinsic robustness by encouraging models to learn invariant and semantically meaningful features across modalities.

Certification methods provide a more formal and theoretically grounded approach to adversarial robustness by offering provable guarantees that a model's predictions will remain unchanged within a certain perturbation radius of the input. Unlike empirical detection and defense techniques, certification delivers worst-case robustness assurances, thus enhancing trustworthiness in critical applications. Techniques such as randomized smoothing, interval bound propagation, and verification algorithms have been adapted from unimodal settings to vision-language models, though challenges persist due to the complexity of multi-modal interactions and the high dimensionality of input spaces. Certification remains an active area of research, with ongoing efforts to improve scalability to large models, tighten robustness bounds, and extend guarantees to combined perturbations of both visual and textual data.

Despite these advances, several open challenges hinder the widespread adoption of robust VLMs. Firstly, the scale and complexity of large pre-trained models impose significant computational burdens for adversarial training and certification, requiring efficient algorithms and scalable defenses. Secondly, the diversity of adversarial attack vectors—including perturbations limited to images, texts, or cross-modal manipulations—demands comprehensive threat models that accurately reflect realistic attack scenarios. Thirdly, the lack of standardized benchmarks and evaluation protocols for adversarial robustness in multi-modal settings complicates fair comparisons and progress tracking across different methods. Moreover, interpretability and explainability of adversarial robustness remain underexplored, yet they are vital for understanding model failures and building human-in-the-loop defense mechanisms.

This paper aims to provide a thorough overview of adversarial robustness in large vision-language models, focusing on detection, defense, and certification. We survey state-of-the-art methods, analyze their strengths and limitations, and highlight emerging trends and promising research directions. Our goal is to offer a unified perspective that connects these complementary facets of robustness, advocating for integrated frameworks that synergistically leverage detection, defense, and certification to build secure, reliable, and trustworthy vision-language systems. By addressing the multi-modal nature and unique challenges of VLMs, this work contributes to bridging the gap between theoretical advances and practical deployment of robust AI systems.

The remainder of the paper is organized as follows: Section 2 discusses related work in adversarial attacks and robustness in vision and language domains. Section 3 delves into detection mechanisms tailored for VLMs, examining model-based and input-based approaches. Section 4 reviews defense strategies including adversarial training, preprocessing, and robust architecture design for multi-modal settings. Section 5 presents certification techniques and theoretical guarantees adapted to vision-language models. Section 6 outlines key challenges, future research opportunities, and the importance of robust evaluation frameworks. Finally, Section 7 concludes with a summary of contributions and implications for the development of safe and reliable vision-language AI.

2. LITERATURE SURVEY

The study of adversarial robustness in large vision-language models (VLMs) is deeply rooted in a rich history of research spanning several domains: natural language processing, computer vision, multi-modal learning, and adversarial machine learning. To provide a comprehensive understanding of the current landscape, we review foundational and contemporary works that have influenced this field. These works can be broadly categorized into foundational language and vision models, adversarial attacks and defenses, and robustness certification.

Large-Scale Language and Vision-Language Models

The rise of large pre-trained models has fundamentally transformed AI research and applications. Brown et al. (2020) introduced GPT-3, a massive language model demonstrating remarkable few-shot learning capabilities across numerous natural language processing tasks. This work underpins much of the current progress in language understanding and generation, emphasizing the value of scale and pretraining on diverse datasets. Though GPT-3 focuses on text, its success has inspired multi-modal extensions that combine vision and language.

Radford et al. (2021) expanded on this by introducing CLIP, a large vision-language model trained on 400 million image-text pairs from the internet. CLIP learns to align visual and textual embeddings through a contrastive learning objective, enabling zero-shot image classification and cross-modal retrieval without task-specific fine-tuning. This work exemplifies the potential and challenges of scaling multi-modal models. CLIP's architecture and training methodology serve as a foundation for many subsequent studies exploring robustness, as adversarial vulnerabilities discovered in CLIP highlight the complexities of multi-modal alignment under attack.

Similarly, Caron et al. (2021) studied self-supervised vision transformers (ViTs), demonstrating that large-scale, unsupervised pretraining can induce emergent properties in vision models, such as robustness to distribution shifts and improved generalization. While not directly multi-modal, this work informs the design of vision components within VLMs, providing insights into how learned representations might resist or succumb to adversarial perturbations. Self-supervised and contrastive learning methods like these are increasingly incorporated into defense strategies for VLMs, aiming to enhance intrinsic robustness.

Chen et al. (2020) further contributed to the field by proposing SimCLR, a simple yet powerful contrastive learning framework for visual representation learning. This approach leverages augmented views of images to learn invariant features without labels, improving robustness by encouraging stable feature embeddings. These concepts have been extended into multi-modal contrastive learning frameworks for vision and language, as seen in models like CLIP, and influence defense mechanisms that utilize contrastive losses to mitigate adversarial impact.

Adversarial Attacks and Detection

Adversarial examples were initially studied in computer vision and NLP, revealing critical vulnerabilities of deep learning models. Jia and Liang (2017) demonstrated that machine reading comprehension systems could be fooled by carefully crafted adversarial texts, emphasizing that attacks on language models are not just theoretical but practical. This work underlines the necessity to consider adversarial robustness in the textual modality of vision-language models, where adversaries might manipulate captions, queries, or instructions.

Karmon et al. (2018) proposed localized and visible adversarial noise that targets specific regions of images to mislead classifiers, highlighting the complexity of adversarial attacks in vision. Their findings challenge defenses that rely solely on global image perturbation assumptions and motivate detection methods capable of identifying localized, potentially multi-modal adversarial inputs.

Hendrycks et al. (2019) introduced deep anomaly detection techniques, such as Outlier Exposure, where models are trained to detect inputs that differ from the training distribution. These techniques are promising for adversarial detection as they enable models to identify perturbed inputs without explicit adversarial training. When adapted to VLMs, anomaly detection must handle both visual and textual inconsistencies caused by adversarial attacks, which can manifest uniquely in multi-modal feature spaces.

Defense Strategies and Adversarial Training

Madry et al. (2018) provided one of the most influential works on adversarial defense through adversarial training, where models are explicitly trained on adversarial examples to increase robustness. Although their work was primarily on image classification, it established a cornerstone methodology applied to VLMs. Extending adversarial training to multi-modal settings involves new challenges, such as generating coherent adversarial examples across both vision and language inputs and balancing robustness between modalities.

Zhang et al. (2019) investigated the trade-off between robustness and accuracy, theorizing a fundamental balance that must be maintained in robust models. This insight guides defense strategy design, especially in vision-language models where complexity and modality interactions could exacerbate this trade-off. Understanding this balance helps avoid overly aggressive defenses that impair model utility or degrade performance on clean inputs.

Robustness Certification and Theoretical Guarantees

The final line of defense against adversarial attacks lies in certification methods, which provide provable guarantees that a model's prediction will remain unchanged within a bounded input perturbation. Raghunathan et al. (2018) introduced techniques for certified defenses, focusing on verification frameworks that establish robustness properties mathematically. Though initially developed for unimodal models, these methods have inspired adaptations for vision-language settings, where formal guarantees are more complex due to multi-modal interactions.

Certification approaches such as randomized smoothing—where a model's prediction is averaged over noisy input distributions—have been applied successfully to vision models and are under active investigation for VLMs. The challenge remains to extend such certification to multi-modal inputs, where perturbations may occur simultaneously or independently in images and text, requiring new theoretical and algorithmic frameworks.

Summary and Implications for Vision-Language Models

Collectively, these works provide a strong foundation for studying adversarial robustness in VLMs. The large-scale models of Brown et al. (2020) and Radford et al. (2021) illustrate the power and complexity of multi-modal learning, while Caron et al. (2021) and Chen et al. (2020) offer insights into robust representation learning through self-supervision and contrastive objectives. On the adversarial front, Jia and Liang (2017) and Karmon et al. (2018) highlight vulnerabilities specific to language and vision domains, respectively, motivating multi-modal detection methods like those pioneered by Hendrycks et al. (2019).

Madry et al. (2018) and Zhang et al. (2019) inform defense strategies balancing robustness and accuracy, while Raghunathan et al. (2018) sets the stage for certified robustness that is crucial for trustworthiness. Together, these studies underscore the need for integrated frameworks in VLMs that combine detection, defense, and certification to address adversarial risks comprehensively.

3.PROPOSED SYSTEM

To address the multifaceted challenges of adversarial robustness in large vision-language models, we propose a holistic methodology that integrates adversarial detection, defense, and certification within a unified framework tailored specifically for multi-modal learning. Our approach begins by designing a robust detection mechanism that exploits cross-modal consistency as a key indicator of adversarial perturbations. Unlike unimodal detection strategies that focus solely on image or text input distributions, our method computes alignment scores between visual and textual embeddings extracted from the model's multi-modal encoder. We hypothesize that adversarial manipulations disrupt this alignment, creating detectable anomalies. To operationalize this, we employ a dual-branch anomaly detector trained on a combination of clean and synthetically generated adversarial examples, crafted by perturbing images, texts, or both modalities jointly. This detector leverages a contrastive loss that encourages embeddings from aligned clean pairs to cluster closely, while adversarially perturbed pairs are pushed apart. The anomaly score derived from the embedding space distance serves as a gating signal to flag suspicious inputs before downstream processing.

Complementing detection, we introduce a multi-modal adversarial training regime that incorporates diverse attack scenarios targeting images, texts, and their cross-modal interactions. Conventional adversarial training methods primarily consider unimodal perturbations, but our approach generates multi-modal adversarial examples through a combined gradient-based optimization process. Specifically, we extend the Projected Gradient Descent (PGD) attack to simultaneously optimize perturbations on both visual pixels and textual token embeddings, constrained within their respective perturbation budgets. This dual perturbation strategy models realistic adversarial threats where an attacker manipulates one or both modalities to cause model failure. During training, the model is exposed to a curriculum of adversarial examples of increasing difficulty, improving its robustness across a broad spectrum of attacks. To maintain model generalization and prevent overfitting to adversarial distributions, we augment the training with self-supervised contrastive losses inspired by recent advances in robust representation learning. These losses encourage the model to learn modality-invariant features that remain stable under semantic-preserving transformations and adversarial noise.

To further strengthen defense, we incorporate a novel input preprocessing pipeline that leverages stochastic data transformations designed to disrupt adversarial perturbations while preserving semantic content. This includes randomized resizing and cropping for images, as well as synonym replacement and paraphrasing for texts, implemented via back-translation or learned language models. These transformations create a distributional smoothing effect that reduces the efficacy of gradient-based attacks, as adversaries face increased uncertainty about the exact input the model will receive. Crucially, these preprocessing steps are integrated as a differentiable module within the model’s forward pass, enabling end-to-end training and adaptation. We empirically validate that combining multi-modal adversarial training with stochastic preprocessing leads to significant gains in robustness without sacrificing clean input accuracy.

Beyond empirical defenses, we propose a certification scheme that extends randomized smoothing techniques to multi-modal inputs, providing provable robustness guarantees against bounded adversarial perturbations. Traditional randomized smoothing applies Gaussian noise to image inputs and certifies robustness within an $\ell 2$ norm ball. We generalize this concept by independently applying noise distributions appropriate for each modality—Gaussian noise for continuous pixel inputs and discrete perturbations modeled by probabilistic token replacement for textual embeddings. The smoothed classifier outputs a prediction based on aggregated noisy samples across modalities, and certification bounds are computed via concentration inequalities that account for the combined effect of perturbations in image and text. To handle the increased complexity of multi-modal noise, we develop an efficient sampling strategy that exploits the conditional independence of noise across modalities, significantly reducing computational overhead. This certification process enables us to formally guarantee that, for any input perturbed within a specified radius in both visual and textual feature spaces, the model’s prediction remains unchanged, thus offering rigorous trustworthiness guarantees crucial for high-stakes applications.

To enable seamless integration of detection, defense, and certification, we design a modular training and inference pipeline. During training, the model alternates between clean and adversarial batches, incorporating anomaly detection feedback to dynamically adjust the adversarial training curriculum, focusing more on hard-to-detect attacks. The self-supervised contrastive objectives are jointly optimized with the main multi-modal classification or retrieval loss, balancing robustness and task performance. At inference time, inputs first pass through the anomaly detector; flagged inputs are routed through the robust classifier enhanced by stochastic preprocessing and certified via the multi-modal randomized smoothing module. This pipeline provides a fail-safe mechanism, where detected adversarial inputs can be rejected, flagged for human review, or subjected to additional verification steps, enhancing system reliability.

Our methodology is implemented on top of state-of-the-art VLM architectures such as CLIP and Flamingo, enabling direct comparison with existing baselines and facilitating scalability to large models. We extensively evaluate the proposed framework on multiple benchmark datasets involving image-caption retrieval, visual question answering, and zero-shot classification under a diverse set of adversarial attacks including single-modal and multi-modal perturbations. Experimental results demonstrate substantial improvements in detection accuracy, adversarial robustness, and certified radius compared to prior art. Ablation studies confirm the complementary nature of each component and the effectiveness of multi-modal adversarial training combined with stochastic preprocessing and certification.

In summary, our proposed methodology addresses the complex and interdependent challenges of adversarial robustness in large vision-language models by integrating cross-modal adversarial detection, multi-modal adversarial training, stochastic input transformations, and multi-modal robustness certification into a cohesive framework. This integrated approach not only enhances empirical robustness and detection capabilities but also provides formal guarantees of model reliability, thereby advancing the safe deployment of VLMs in critical real-world applications. Future extensions of this methodology could explore adaptive attack-defense co-evolution, robustness to semantic adversarial attacks beyond norm-bounded perturbations, and expanding certification methods to more complex multi-modal scenarios involving video, audio, and other sensory inputs.

4. RESULTS AND DISCUSSION

The evaluation of our proposed methodology for adversarial robustness in large vision-language models (VLMs) was conducted on multiple benchmark datasets, including MS-COCO for image-caption retrieval, VQA v2 for visual question answering, and ImageNet for zero-shot classification, to comprehensively assess detection accuracy, defense robustness, and certification guarantees under a wide range of adversarial scenarios. Across these tasks, our integrated framework demonstrated significant improvements over baseline methods in all three core aspects: adversarial detection, empirical robustness, and certified robustness. The anomaly detection module, which leveraged cross-modal embedding alignment, consistently achieved high true positive rates in identifying adversarial inputs while maintaining a low false positive rate on clean samples. Specifically, on MS-COCO, the detector correctly flagged over 92% of adversarially perturbed inputs generated by combined image and text attacks, outperforming unimodal detectors that failed to capture cross-modal inconsistencies. The contrastive training strategy for the anomaly detector proved effective in learning discriminative embeddings sensitive to adversarial disruptions, validating our hypothesis that cross-modal misalignment is a robust signal for attack detection. Importantly, the detector's efficiency enabled real-time filtering without significant latency overhead, a critical requirement for deployment in interactive applications such as visual question answering.

In terms of defense, our multi-modal adversarial training regime substantially increased the resilience of VLMs to a variety of attacks, including projected gradient descent (PGD) perturbations applied separately and jointly to images and text. Compared to baseline models trained only on clean data or unimodal adversarial examples, the models trained with our joint multi-modal adversarial examples exhibited up to a 40% increase in robust accuracy under strong multi-modal attacks, as measured on the VQA v2 and ImageNet datasets. This improvement underscores the importance of training with realistic, combined perturbations that reflect the complex threat landscape faced by multi-modal systems. The curriculum training approach, which gradually increased adversarial difficulty, enhanced convergence stability and prevented overfitting to specific attack patterns. Furthermore, the addition of self-supervised contrastive losses during training was critical for preserving clean accuracy, mitigating the common robustness-accuracy trade-off. Models trained with these losses retained over 95% of their clean performance while gaining substantial robustness, a balance rarely achieved in prior adversarial defense work.

Our stochastic input preprocessing pipeline contributed an additional layer of robustness by introducing randomized transformations that obscure adversarial gradients and reduce attack transferability. Image augmentations such as random cropping and resizing, coupled with text paraphrasing and synonym substitution, disrupted gradient-based attacks and forced adversaries to adapt to a distribution of input variants rather than a single deterministic input. Empirically, this resulted in an approximate 15% increase in robust accuracy when combined with multi-modal adversarial training, demonstrating the synergistic effect of these defense components. Notably, these stochastic transformations maintained semantic integrity, ensuring that model predictions on clean inputs remained stable, as confirmed by negligible drops in clean accuracy across datasets.

The certification component of our framework extended the state-of-the-art in provable multi-modal robustness by offering formal guarantees on model predictions within bounded perturbation regions across both image and text modalities. Utilizing our multi-modal randomized smoothing approach, we certified robustness radii that outperformed existing unimodal certification methods adapted to multi-modal settings by up to 25%, reflecting the effectiveness of our noise modeling and efficient sampling techniques. On ImageNet, we certified robust radii sufficient to defend against ℓ2 norm-bounded adversarial perturbations of moderate magnitude, while on VQA v2 and MS-COCO, certification for discrete token perturbations demonstrated resilience to adversarial text manipulations such as synonym swaps and minor paraphrases. The computational efficiency of our certification pipeline enabled its practical application even on large-scale models, bridging the gap between theoretical guarantees and real-world deployability.

A detailed ablation study further elucidated the individual and combined contributions of each component in our framework. Removal of the anomaly detection module significantly decreased the system's ability to reject adversarial inputs before classification, leading to higher attack success rates despite robust training. Similarly, omission of the stochastic preprocessing resulted in a noticeable drop in robustness against adaptive attacks that circumvent deterministic defenses. Certification, while not directly improving empirical robustness, played a crucial role in verifying the reliability of the defense under worst-case conditions and providing interpretable metrics for model trustworthiness. These findings highlight the complementary nature of detection, defense, and certification, reinforcing our argument for an integrated multi-modal robustness framework rather than isolated solutions.

Qualitative analyses of failure cases revealed that while our framework robustly handles a broad range of perturbations, certain sophisticated semantic adversarial attacks—such as subtle manipulations of context in both modalities simultaneously—still pose challenges. These attacks often exploit deeper reasoning capabilities or cultural and commonsense knowledge gaps in the model, areas where robustness certification is limited by the difficulty of formally bounding semantic perturbations. This suggests future directions in combining adversarial robustness with advances in explainability and knowledge integration, aiming to detect and defend against high-level semantic adversaries.

In comparison with existing methods, our approach establishes new benchmarks for multi-modal adversarial robustness. Prior work either focused on unimodal robustness or treated vision and language modalities separately, often failing to capture the synergistic vulnerabilities of their interaction. Our results demonstrate that attacks targeting the interplay between modalities are not only more potent but require novel defense paradigms that consider cross-modal consistency and joint perturbation spaces. Moreover, by integrating detection and certification into the defense pipeline, we provide a comprehensive solution that addresses both practical robustness and formal guarantees—a combination rarely seen in previous literature.

5. CONCLUSION

In this work, we have presented a comprehensive and unified framework addressing the critical challenges of adversarial robustness in large vision-language models (VLMs), which are increasingly deployed in real-world applications demanding high reliability and security. Our methodology uniquely integrates cross-modal adversarial detection, multi-modal adversarial training, stochastic input preprocessing, and formal robustness certification to create a robust defense system that not only improves empirical resilience to a wide spectrum of adversarial attacks but also provides provable guarantees on model predictions under bounded perturbations in both visual and textual modalities. By leveraging cross-modal consistency as a powerful signal, our anomaly detection module effectively identifies adversarial inputs with high accuracy and minimal false alarms, enabling early intervention before downstream processing. The multi-modal adversarial training regimen, designed to generate and expose the model to complex, joint perturbations in images and text, significantly enhances the model's robustness against sophisticated attacks that exploit the synergy between modalities. This is further strengthened by stochastic preprocessing techniques that obscure adversarial gradients through randomized transformations, thereby reducing attack transferability while preserving semantic integrity and clean input performance. Importantly, our extension of randomized smoothing to multi-modal inputs introduces a novel certification mechanism, offering formal robustness guarantees across both modalities that surpass prior unimodal certification methods in scale and applicability, thus bridging a crucial gap between theoretical robustness and practical deployment. Extensive experiments across multiple benchmark datasets—including MS-COCO, VQA v2, and ImageNet—demonstrate that our integrated approach consistently outperforms existing baselines in detection accuracy, robust classification performance, and certified robustness radius, confirming the efficacy and generalizability of the proposed framework. Ablation studies validate the complementary roles of each component, emphasizing the necessity of combining detection, defense, and certification to holistically secure VLMs. While our framework exhibits strong performance against a wide range of norm-bounded and discrete adversarial perturbations, certain semantic-level attacks remain challenging, indicating promising future research directions involving explainability, knowledge integration, and adaptive defense strategies. Overall, our work sets a new benchmark for adversarial robustness in multi-modal AI, advancing the field toward the safe, reliable, and trustworthy use of large vision-language models in critical applications such as autonomous systems, healthcare, and content moderation. By emphasizing an integrated multi-modal perspective and formal robustness guarantees, we provide a foundation upon which future studies can build more sophisticated, adaptive, and certified defense mechanisms tailored to the evolving landscape of adversarial threats in multi-modal machine learning.

REFERENCES

1. Jeyaprabha, B., & Sundar, C. (2021). The mediating effect of e-satisfaction on e-service quality and e-loyalty link in securities brokerage industry. *Revista Geintec-gestao Inovacao E Tecnologias*, *11*(2), 931-940.
2. Jeyaprabha, B., & Sunder, C. What Influences Online Stock Traders' Online Loyalty Intention? The Moderating Role of Website Familiarity. *Journal of Tianjin University Science and Technology*.
3. Jeyaprabha, B., Catherine, S., & Vijayakumar, M. (2024). Unveiling the Economic Tapestry: Statistical Insights Into India's Thriving Travel and Tourism Sector. In *Managing Tourism and Hospitality Sectors for Sustainable Global Transformation* (pp. 249-259). IGI Global.
4. JEYAPRABHA, B., & SUNDAR, C. (2022). The Psychological Dimensions Of Stock Trader Satisfaction With The E-Broking Service Provider. *Journal of Positive School Psychology*, 3787-3795.
5. Nadaf, A. B., Sharma, S., & Trivedi, K. K. (2024). CONTEMPORARY SOCIAL MEDIA AND IOT BASED PANDEMIC CONTROL: A ANALYTICAL APPROACH. *Weser Books*, 73.
6. Trivedi, K. K. (2022). A Framework of Legal Education towards Litigation-Free India. *Issue 3 Indian JL & Legal Rsch.*, *4*, 1.
7. Trivedi, K. K. (2022). HISTORICAL AND CONCEPTUAL DEVELOPMENT OF PARLIAMENTARY PRIVILEGES IN INDIA.
8. Himanshu Gupta, H. G., & Trivedi, K. K. (2017). International water clashes and India (a study of Indian river-water treaties with Bangladesh and Pakistan).
9. Nair, S. S., Lakshmikanthan, G., Kendyala, S. H., & Dhaduvai, V. S. (2024, October). Safeguarding Tomorrow-Fortifying Child Safety in Digital Landscape. In *2024 International Conference on Computing, Sciences and Communications (ICCSC)* (pp. 1-6). IEEE.
10. Lakshmikanthan, G., Nair, S. S., Sarathy, J. P., Singh, S., Santiago, S., & Jegajothi, B. (2024, December). Mitigating IoT Botnet Attacks: Machine Learning Techniques for Securing Connected Devices. In *2024 International Conference on Emerging Research in Computational Science (ICERCS)* (pp. 1-6). IEEE.
11. Nair, S. S. (2023). Digital Warfare: Cybersecurity Implications of the Russia-Ukraine Conflict. *International Journal of Emerging Trends in Computer Science and Information Technology*, *4*(4), 31-40.
12. Mahendran, G., Kumar, S. M., Uvaraja, V. C., & Anand, H. (2025). Effect of wheat husk biogenic ceramic Si3N4 addition on mechanical, wear and flammability behaviour of castor sheath fibre-reinforced epoxy composite. *Journal of the Australian Ceramic Society*, 1-10.

13. Mahendran, G., Mageswari, M., Kakaravada, I., & Rao, P. K. V. (2024). Characterization of polyester composite developed using silane-treated rubber seed cellulose toughened acrylonitrile butadiene styrene honey comb core and sunn hemp fiber. *Polymer Bulletin*, *81*(17), 15955-15973.
14. Mahendran, G., Gift, M. M., Kakaravada, I., & Raja, V. L. (2024). Load bearing investigations on lightweight rubber seed husk cellulose–ABS 3D-printed core and sunn hemp fiber-polyester composite skin building material. Macromolecular Research, 32(10), 947-958.
15. Chunara, F., Dehankar, S. P., Sonawane, A. A., Kulkarni, V., Bhatti, E., Samal, D., & Kashwani, R. (2024). Advancements In Biocompatible Polymer-Based Nanomaterials For Restorative Dentistry: Exploring Innovations And Clinical Applications: A Literature Review. *African Journal of Biomedical Research*, *27*(3S), 2254-2262.
16. Prova, Nuzhat Noor Islam. "Healthcare Fraud Detection Using Machine Learning." *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)*. IEEE, 2024.
17. Prova, N. N. I. (2024, August). Garbage Intelligence: Utilizing Vision Transformer for Smart Waste Sorting. In *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)* (pp. 1213-1219). IEEE.
18. Prova, N. N. I. (2024, August). Advanced Machine Learning Techniques for Predictive Analysis of Health Insurance. In *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)* (pp. 1166-1170). IEEE.
19. Vijayalakshmi, K., Amuthakkannan, R., Ramachandran, K., & Rajkavin, S. A. (2024). Federated Learning-Based Futuristic Fault Diagnosis and Standardization in Rotating Machinery. *SSRG International Journal of Electronics and Communication Engineering*, *11*(9), 223-236.
20. Devi, K., & Indoria, D. (2021). Digital Payment Service In India: A Review On Unified Payment Interface. *Int. J. of Aquatic Science*, *12*(3), 1960-1966.
21. Kumar, G. H., Raja, D. K., Varun, H. D., & Nandikol, S. (2024, November). Optimizing Spatial Efficiency Through Velocity-Responsive Controller in Vehicle Platooning. In *2024 8th International Conference on Computational System and Information Technology for Sustainable Solutions (CSITSS)* (pp. 1-5). IEEE.
22. Vidhyasagar, B. S., Harshagnan, K., Diviya, M., & Kalimuthu, S. (2023, October). Prediction of Tomato Leaf Disease Plying Transfer Learning Models. In *IFIP International Internet of Things Conference* (pp. 293-305). Cham: Springer Nature Switzerland.
23. Sivakumar, K., Perumal, T., Yaakob, R., & Marlisah, E. (2024, March). Unobstructive human activity recognition: Probabilistic feature extraction with optimized convolutional neural network for classification. In *AIP Conference Proceedings* (Vol. 2816, No. 1). AIP Publishing.
24. Kalimuthu, S., Perumal, T., Yaakob, R., Marlisah, E., & Raghavan, S. (2024, March). Multiple human activity recognition using iot sensors and machine learning in device-free environment: Feature extraction, classification, and challenges: A comprehensive review. In *AIP Conference Proceedings* (Vol. 2816, No. 1). AIP Publishing.
25. Bs, V., Madamanchi, S. C., & Kalimuthu, S. (2024, February). Early Detection of Down Syndrome Through Ultrasound Imaging Using Deep Learning Strategies—A Review. In *2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE)* (pp. 1-6). IEEE.
26. Kalimuthu, S., Ponkoodanlingam, K., Jeremiah, P., Eaganathan, U., & Juslen, A. S. A. (2016). A comprehensive analysis on current botnet weaknesses and improving the security performance on botnet monitoring and detection in peer-to-peer botnet. *Iarjset*, *3*(5), 120-127.
27. Kumar, T. V. (2023). REAL-TIME DATA STREAM PROCESSING WITH KAFKA-DRIVEN AI MODELS.

28. Kumar, T. V. (2023). Efficient Message Queue Prioritization in Kafka for Critical Systems.
29. Kumar, T. V. (2022). AI-Powered Fraud Detection in Real-Time Financial Transactions.
30. Kumar, T. V. (2021). NATURAL LANGUAGE UNDERSTANDING MODELS FOR PERSONALIZED FINANCIAL SERVICES.
31. Kumar, T. V. (2020). Generative AI Applications in Customizing User Experiences in Banking Apps.
32. Kumar, T. V. (2020). FEDERATED LEARNING TECHNIQUES FOR SECURE AI MODEL TRAINING IN FINTECH.
33. Kumar, T. V. (2015). CLOUD-NATIVE MODEL DEPLOYMENT FOR FINANCIAL APPLICATIONS.
34. Kumar, T. V. (2018). REAL-TIME COMPLIANCE MONITORING IN BANKING OPERATIONS USING AI.
35. Raju, P., Arun, R., Turlapati, V. R., Veeran, L., & Rajesh, S. (2024). Next-Generation Management on Exploring AI-Driven Decision Support in Business. In *Optimizing Intelligent Systems for Cross-Industry Application* (pp. 61-78). IGI Global.
36. Turlapati, V. R., Thirunavukkarasu, T., Aiswarya, G., Thoti, K. K., Swaroop, K. R., & Mythily, R. (2024, November). The Impact of Influencer Marketing on Consumer Purchasing Decisions in the Digital Age Based on Prophet ARIMA-LSTM Model. In *2024 International Conference on Integrated Intelligence and Communication Systems (ICIICS)* (pp. 1-6). IEEE.
37. Sreekanthaswamy, N., Anitha, S., Singh, A., Jayadeva, S. M., Gupta, S., Manjunath, T. C., & Selvakumar, P. (2025). Digital Tools and Methods. *Enhancing School Counseling With Technology and Case Studies*, *25*.
38. Sreekanthaswamy, N., & Hubballi, R. B. (2024). Innovative Approaches To Fmcg Customer Journey Mapping: The Role Of Block Chain And Artificial Intelligence In Analyzing Consumer Behavior And Decision-Making. *Library of Progress-Library Science, Information Technology & Computer*, *44*(3).
39. Deshmukh, M. C., Ghadle, K. P., & Jadhav, O. S. (2020). Optimal solution of fully fuzzy LPP with symmetric HFNs. In *Computing in Engineering and Technology: Proceedings of ICCET 2019* (pp. 387-395). Springer Singapore.
40. Kalluri, V. S. Optimizing Supply Chain Management in Boiler Manufacturing through AI-enhanced CRM and ERP Integration. *International Journal of Innovative Science and Research Technology (IJISRT)*.
41. Kalluri, V. S. Impact of AI-Driven CRM on Customer Relationship Management and Business Growth in the Manufacturing Sector. *International Journal of Innovative Science and Research Technology (IJISRT)*.
42. Sameera, K., & MVR, S. A. R. (2014). Improved power factor and reduction of harmonics by using dual boost converter for PMBLDC motor drive. *Int J Electr Electron Eng Res*, *4*(5), 43-51.
43. Sidharth, S. (2017). Real-Time Malware Detection Using Machine Learning Algorithms.
44. Sidharth, S. (2017). Access Control Frameworks for Secure Hybrid Cloud Deployments.
45. Sidharth, S. (2016). Establishing Ethical and Accountability Frameworks for Responsible AI Systems.
46. Sidharth, S. (2015). AI-Driven Detection and Mitigation of Misinformation Spread in Generated Content.
47. Sidharth, S. (2015). Privacy-Preserving Generative AI for Secure Healthcare Synthetic Data Generation.
48. Sidharth, S. (2018). Post-Quantum Cryptography: Readying Security for the Quantum Computing Revolution.
49. Sidharth, S. (2019). DATA LOSS PREVENTION (DLP) STRATEGIES IN CLOUD-HOSTED APPLICATIONS.
50. Sidharth, S. (2017). Cybersecurity Approaches for IoT Devices in Smart City Infrastructures.