Communications, Electronics and Sensing Systems
https://ces-systems.org

Adversarial Robustness Evaluation for Vision Foundation Models and Generative Artificial Intelligence Systems

Eun-Ji Kang, Soo-Min Song

Department of Electrical Engineering, School of Electronic and Electrical Engineering, Kyungpook National University, Daegu 41566, Korea

Editorial correspondence on behalf of the authors: info@ces-systems.org

Abstract

The rapid advancement of artificial intelligence has ushered in a new era dominated by vision foundation models and generative artificial intelligence systems. These architectures exhibit unprecedented capabilities in image synthesis, classification, and cross-modal understanding. However, their increasing deployment in safety-critical domains raises substantial concerns regarding their susceptibility to adversarial perturbations. This paper presents a comprehensive evaluation of adversarial robustness specifically tailored to the unique vulnerabilities of large-scale vision and generative models. By systematically analyzing the attack surfaces, including input space perturbations and prompt injection techniques, we illuminate the complex failure modes inherent in these advanced systems. Our investigation encompasses a rigorous methodological framework that assesses the degradation of model performance under various threat models, ranging from white-box gradient-based attacks to black-box query-based methodologies. The extensive analysis demonstrates that while foundation models possess remarkable generalization capabilities, they simultaneously inherit and amplify adversarial vulnerabilities from their pre-training corpora. Furthermore, we explore the intricate balance between generative fidelity and adversarial resilience, revealing significant trade-offs that complicate the development of robust defensive mechanisms. Through empirical investigation and theoretical analysis, this research provides critical insights into the limitations of current generative architectures and proposes foundational principles for the design of next-generation, robust artificial intelligence systems.

Keywords: Vision Foundation Models; Generative Artificial Intelligence; Adversarial Robustness; Threat Models

1. Introduction

1.1 Context and Motivation

The landscape of artificial intelligence has undergone a fundamental transformation over the past decade, evolving from task-specific algorithms to highly generalized foundation models. These systems, characterized by their massive scale and training on vast amounts of uncurated data, have demonstrated remarkable capabilities across a diverse array of applications. In the realm of computer vision, foundation models have fundamentally altered how machines perceive and interpret visual data [1]. Unlike earlier convolutional neural networks that were explicitly trained for narrow classification or detection tasks, modern vision foundation models leverage cross-modal self-supervision to develop a comprehensive understanding of the semantic relationships between images and natural language. This paradigm shift has enabled zero-shot learning capabilities that were previously unattainable, allowing systems to generalize to novel concepts without the need for extensive task-specific fine-tuning [2]. Concurrently, the emergence of generative artificial intelligence systems, particularly those based on diffusion processes, has revolutionized digital content creation. These generative models can synthesize high-fidelity, photorealistic images from textual descriptions, offering unprecedented creative control and utility in fields ranging from entertainment to industrial design [3].

However, the widespread integration of these powerful models into societal infrastructure necessitates a rigorous examination of their reliability and security. As foundation models become the underlying engines for autonomous navigation systems, automated medical diagnostics, and automated content moderation, the consequences of system failure become increasingly severe [4]. The core motivation for this research stems from the well-documented susceptibility of deep neural networks to adversarial examples, which are carefully crafted input perturbations designed to deceive machine learning models. While adversarial vulnerability is a known flaw in traditional architectures, the manifestation of these vulnerabilities in contemporary foundation and generative models introduces novel and highly complex security challenges [5]. The vast parameter spaces and complex internal representations of these modern systems may inadvertently create new attack vectors that malicious actors could exploit. Consequently, understanding and mitigating these adversarial threats is not merely an academic exercise but a critical prerequisite for the safe and responsible deployment of next-generation artificial intelligence.

1.2 Challenges in Robustness Evaluation

Evaluating the adversarial robustness of vision foundation models and generative systems presents a unique set of methodological and computational challenges that depart significantly from traditional robustness assessments. In conventional image classification, an adversarial attack is typically considered successful if a visually imperceptible perturbation causes the model to output an incorrect class label. The evaluation metric is straightforward, relying on a binary assessment of classification accuracy under attack [6]. However, the objectives and outputs of foundation and

generative models are multifaceted, rendering traditional evaluation paradigms insufficient. For cross-modal vision-language models, an attacker might seek to disrupt the alignment between an image and its corresponding text, thereby degrading the model's retrieval capabilities or forcing it to associate an image with a completely unrelated semantic concept [7]. Measuring the success of such an attack requires nuanced metrics that quantify semantic drift and representational distortion.

The evaluation of generative systems introduces even greater complexity. When assessing a text-to-image generation model, an adversarial attack might aim to manipulate the input text prompt to bypass safety filters, or it might target an input conditioning image to force the generation of specific, unintended artifacts [8]. Determining the success of an attack on a generative model involves subjective and objective assessments of image quality, fidelity to the prompt, and the presence of malicious structural deviations. Furthermore, the sheer scale of these models imposes significant computational constraints on robustness evaluation. Gradient-based attack methodologies, which necessitate multiple backward passes through the network to iteratively refine perturbations, become computationally prohibitive when applied to models containing billions of parameters [9]. Consequently, researchers must develop highly efficient optimization strategies and evaluation frameworks capable of scaling to contemporary architectures without sacrificing the rigor of the security assessment [10]. This paper directly addresses these challenges by formulating a comprehensive evaluation methodology tailored to the unique operational characteristics of these advanced systems.

2. Literature Review and Background

2.1 Evolution of Vision Foundation Models

The historical trajectory of computer vision is marked by significant architectural innovations that have progressively enhanced the capacity of machines to process visual information. Early deep learning approaches were dominated by convolutional neural networks, which leveraged local receptive fields and shared weights to efficiently extract hierarchical features from images [11]. While highly successful on constrained benchmarks, these architectures often struggled with global context and required immense amounts of meticulously labeled data to achieve high performance on novel tasks. The introduction of the transformer architecture, initially designed for natural language processing, catalyzed a major paradigm shift. By processing images as sequences of patches and employing self-attention mechanisms, vision transformers demonstrated an unprecedented ability to capture long-range dependencies and global structural relationships within visual data [12].

This architectural evolution laid the groundwork for the development of vision foundation models, which seek to learn highly generalized representations applicable to a wide variety of downstream tasks. A seminal advancement in this domain was the development of contrastive language-image pre-training methodologies. By training models to maximize the similarity between corresponding images and text descriptions within a shared latent space, these systems learned to associate visual features with rich semantic concepts [13]. This cross-modal alignment proved highly effective, enabling capabilities such as zero-shot classification, where a model can accurately categorize images based solely on textual prompts defining the categories, without ever having been explicitly trained on those specific labels. The success of contrastive learning relies heavily on the availability of massive, internet-scale datasets comprising billions of image-text pairs [14]. However, the uncurated nature of these pre-training corpora introduces significant risks, as the models inevitably absorb the biases, noise, and potential adversarial vulnerabilities present within the training data, necessitating rigorous downstream robustness evaluation.

2.2 Generative Artificial Intelligence Systems

Parallel to the advancements in discriminative vision models, generative artificial intelligence has experienced a period of explosive growth, driven by fundamental breakthroughs in generative modeling paradigms. Early approaches, such as generative adversarial networks, framed the image synthesis process as a competitive game between a generator attempting to create realistic data and a discriminator attempting to distinguish real from synthetic data [15]. While capable of producing high-quality images, generative adversarial networks frequently suffered from training instability and mode collapse, where the generator would learn to produce only a limited subset of possible outputs. The field subsequently gravitated towards diffusion models, which offer a more stable and mathematically grounded approach to generative modeling.

Diffusion models operate by defining a forward process that gradually adds Gaussian noise to an image until it is reduced to pure random noise, followed by a reverse process where a neural network is trained to iteratively denoise the data, progressively reconstructing the original image [16]. By conditioning this reverse denoising process on textual embeddings derived from powerful language models, researchers created highly controllable text-to-image synthesis pipelines. The integration of latent space operations further enhanced the efficiency of diffusion models, allowing the computationally intensive denoising steps to occur in a compressed representational space rather than the high-dimensional pixel space [17]. This innovation democratized access to high-fidelity generative systems, enabling rapid synthesis of complex scenes on standard computational hardware. Despite these remarkable achievements, the iterative nature of the diffusion process and its reliance on external conditioning mechanisms introduce unique attack surfaces. An adversary might subtly manipulate the initial noise distribution, the conditioning text, or intermediate latent representations to drastically alter the final generated output, highlighting the critical need for comprehensive vulnerability assessments [18].

2.3 Adversarial Attack Paradigms

The study of adversarial machine learning has established a robust theoretical framework for understanding the vulnerabilities of deep neural networks. At its core, an adversarial attack involves the intentional modification of input data to cause a machine learning system to fail, typically while ensuring that the modification remains imperceptible or innocuous to a human observer [19]. Attacks are generally categorized based on the adversary's knowledge of the target system. In a white-box scenario, the attacker possesses complete knowledge of the model's architecture, parameters, and internal state, allowing for the precise computation of gradients to guide the perturbation process [20]. Conversely, a black-box scenario assumes the attacker has no internal knowledge of the model and can only interact with it by providing inputs and observing the corresponding outputs.

Early white-box attack methodologies focused on efficiency, utilizing single-step gradient updates to quickly generate perturbations that maximize the model's classification error. Subsequent research demonstrated that iterative optimization techniques, which apply multiple small gradient steps while projecting the perturbed input back into a valid constraint bound, produce significantly stronger and more reliable adversarial examples [21]. In the context of vision foundation models and generative systems, these traditional paradigms must be adapted. Attackers targeting cross-modal models often seek to maximize the distance between an image and its corresponding text embedding, rather than simply altering a discrete class label [22]. For generative models, adversarial objectives become even more complex, encompassing goals such as forcing the generation of highly specific, targeted visual artifacts or entirely subverting the semantic meaning of the input prompt [23]. Furthermore, the phenomenon of adversarial transferability, where perturbations designed for one model successfully deceive a different, unseen model, poses a severe threat in the foundation model era, as attackers can generate transferable perturbations using accessible surrogate models to attack proprietary, black-box systems deployed in production environments.

3. Methodology for Robustness Evaluation

3.1 Threat Model Definition

To systematically evaluate the adversarial robustness of vision foundation models and generative systems, it is imperative to establish rigorously defined threat models that encapsulate the diverse capabilities and objectives of potential adversaries. A threat model formally specifies the constraints under which an attacker operates, providing a standardized framework for assessing the relative vulnerability of different architectures. In this research, we articulate threat models across three primary dimensions encompassing knowledge, capability, and objective.

Regarding adversarial knowledge, we evaluate models under both white-box and black-box conditions. The white-box threat model assumes the adversary has unfettered access to the target model's internal architecture, weight parameters, and gradients [24]. This scenario represents the absolute worst-case security posture and is crucial for determining the theoretical lower bound of a model's robustness. In contrast, the black-box threat model restricts the adversary to querying the model and analyzing its outputs. We simulate black-box conditions through transferability assessments, where attacks are generated on a known local substitute model and subsequently applied to the target system.

Regarding adversarial capability, we focus primarily on test-time evasion attacks, where the adversary introduces imperceptible perturbations to the input data during the inference phase [25]. We constrain the magnitude of these perturbations using standard distance metrics to ensure they remain visually undetectable to human observers. For text-conditioned generative models, we expand the capability dimension to include prompt injection techniques, where the adversary carefully crafts malicious input text designed to override the system's intended behavior and bypass safety guardrails.

Regarding adversarial objectives, we distinguish between untargeted and targeted attacks. An untargeted attack on a vision foundation model aims simply to disrupt the alignment between the image and its true semantic representation, causing general system failure. A targeted attack, however, seeks to force the model to associate the input image with a specific, adversarial chosen semantic concept [26]. For generative systems, targeted attacks attempt to induce the synthesis of predefined visual patterns or objects regardless of the user's intended prompt, demonstrating a severe compromise of generative control.

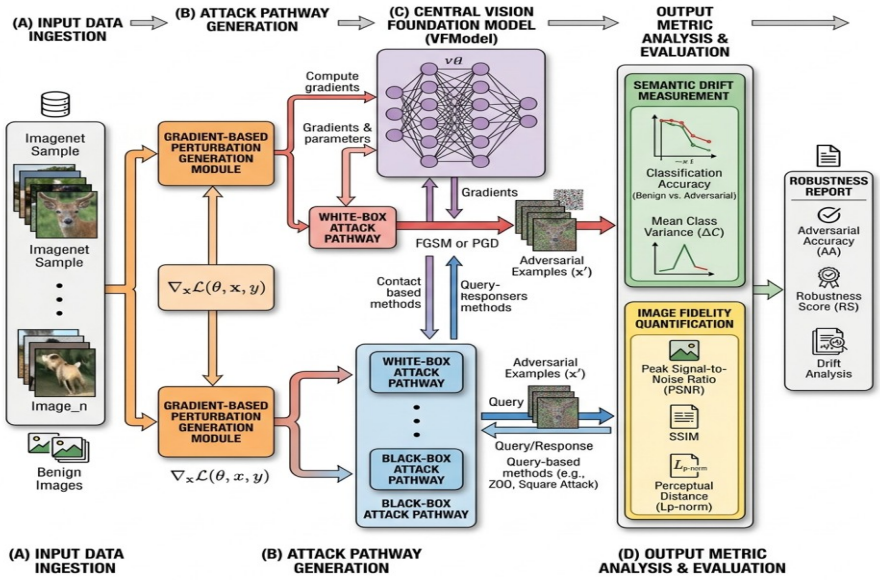

Figure 1: Comprehensive Architectural Schematic of the Adversarial Robustness Evaluation Pipeline

3.2 Perturbation Generation Strategies

The process of generating effective adversarial perturbations requires sophisticated optimization strategies tailored to the specific architecture and objective of the target model. For vision foundation models that rely on contrastive learning, the primary goal of the perturbation generation strategy is to manipulate the latent embeddings produced by the image encoder. We employ iterative optimization techniques that leverage the gradients of the model's loss function with respect to the input pixels.

In our evaluation framework, the perturbation process begins by selecting an original, unperturbed image and its corresponding textual description. We define an adversarial loss function designed to maximize the distance between the image embedding and the text embedding in the shared latent space [27]. For a targeted attack, the loss function is modified to simultaneously minimize the distance between the perturbed image embedding and a predefined target text embedding. We initialize a small perturbation tensor, which is added to the original image. During each iteration of the optimization process, we perform a forward pass through the image encoder to compute the current loss, followed by a backward pass to calculate the gradient of the loss with respect to the input image. The perturbation is then updated in the direction of the gradient to maximize the adversarial objective. To ensure the perturbation remains imperceptible, we apply a projection function after each update, constraining the magnitude of the perturbation to a specified bound.

For generative artificial intelligence systems based on diffusion processes, the perturbation generation strategy must account for the iterative nature of the denoising mechanism. Attacks can be directed at either the input conditioning image used for tasks like image-to-image translation, or the textual prompt used for text-to-image synthesis. When attacking the input image, the optimization objective is to find a perturbation that minimizes the difference between the final generated image and a desired adversarial target image, necessitating the computation of gradients through the entire reverse diffusion process [28]. Due to the massive computational overhead of this approach, we also employ approximation techniques that target the latent representations at specific timesteps within the diffusion process, effectively hijacking the generation trajectory before the final image is fully realized.

3.3 Evaluation Metrics and Framework

Establishing robust and comprehensive evaluation metrics is critical for accurately quantifying the impact of adversarial perturbations on vision foundation models and generative systems. Due to the diverse functionalities of these models, a multi-faceted metric framework is required. For vision foundation models evaluated on retrieval and zero-shot classification tasks, we utilize standard performance degradation metrics. The primary indicator of vulnerability is the attack success rate, which measures the percentage of adversarial examples that successfully induce an incorrect output or semantic misalignment [29]. We also evaluate the drop in top-1 and top-5 accuracy under various perturbation budgets to understand the model's degradation curve.

In the context of generative artificial intelligence, traditional classification metrics are inapplicable. Instead, we must quantify both the visual quality of the generated output and its adherence to the intended semantics. To measure the distortion introduced by the attack, we employ perceptual distance metrics that compare the structural and textural similarities between images generated from clean inputs and those generated from perturbed inputs. A successful attack on a generative model often results in significant structural deviations, which are captured by these distance measurements [30]. Furthermore, we assess the semantic fidelity of the generated images using automated scoring mechanisms that evaluate how well the output image aligns with the original text prompt. An effective adversarial attack will drastically

reduce this alignment score, indicating that the generative model has been forced to synthesize content irrelevant to the user's request.

Our comprehensive evaluation framework integrates these diverse metrics into a unified pipeline. The framework systematically ingests benchmark datasets, applies the defined perturbation generation strategies across multiple perturbation constraints, queries the target models, and aggregates the resulting metric data. This automated pipeline ensures reproducibility and allows for large-scale comparative analysis across different foundation model architectures and generative paradigms, providing a rigorous empirical basis for our vulnerability assessments.

4. Experimental Design and Setup

4.1 Dataset Selection and Preparation

The empirical evaluation of adversarial robustness necessitates the careful selection of datasets that provide a diverse and representative sample of visual and semantic concepts. For our experimental design, we utilize subsets derived from established large-scale computer vision benchmarks to ensure broad coverage and relevance. To evaluate the zero-shot classification capabilities of vision foundation models, we construct a high-quality evaluation set sampled from a widely recognized image classification database. This database contains high-resolution images categorized into thousands of distinct classes, providing a rigorous test bed for assessing the model's ability to maintain semantic alignment under adversarial conditions.

For evaluating the robustness of cross-modal alignment and generative artificial intelligence systems, we require datasets that provide rich, descriptive text pairings for each image. We curate a specialized evaluation corpus from massive internet-scraped datasets, specifically selecting image-text pairs that exhibit high mutual information and descriptive complexity. The preparation of this data involves stringent filtering processes to remove instances with ambiguous text descriptions, highly compressed images, or inappropriate content that could skew the robustness evaluation [31]. All selected images are centrally cropped and resized to a standardized resolution dictated by the input requirements of the target models, ensuring uniformity across all experimental trials. Furthermore, we normalize the pixel values of the dataset to match the specific statistical distributions expected by the respective pre-trained models, a crucial step for maintaining the validity of gradient-based perturbation techniques.

4.2 Target Models and Configurations

Our experimental setup is designed to evaluate a representative cross-section of contemporary state-of-the-art vision foundation models and generative architectures. We select target models that demonstrate distinct structural paradigms to assess how different architectural choices influence adversarial vulnerability. The first category of target models encompasses contrastive language-image pre-training systems utilizing transformer-based image encoders. We select variants with different parameter counts and patch sizes to investigate the relationship between model scale and robustness. These models serve as the primary subjects for evaluating cross-modal semantic disruption.

The second category comprises advanced generative artificial intelligence systems, specifically latent diffusion models. We evaluate popular open-weight text-to-image architectures that utilize cross-attention mechanisms to condition the generative process on text embeddings. These systems are evaluated on their susceptibility to both input image perturbations and prompt injection techniques. The configuration of these target models is meticulously controlled to ensure reproducible results. All models are initialized with their official, publicly available pre-trained weights, representing the exact configurations deployed in numerous real-world applications. During the evaluation phase, the models are operated strictly in inference mode, with all trainable parameters frozen. For the diffusion models, we standardize the generation process by fixing the number of denoising steps and employing a consistent deterministic sampler, thereby isolating the effects of the adversarial perturbations from the inherent randomness of the generative process.

4.3 Implementation Constraints

Executing adversarial robustness evaluations on billion-parameter foundation models imposes formidable computational demands, requiring specific implementation constraints and hardware optimizations. The iterative calculation of gradients through massive transformer blocks and multi-step diffusion processes consumes substantial amounts of memory and processing time. To mitigate these bottlenecks, all experiments are conducted on high-performance computational clusters equipped with advanced tensor processing hardware. We implement gradient checkpointing techniques during the perturbation generation phase, which significantly reduces the peak memory footprint by selectively recomputing intermediate activations during the backward pass rather than storing them in memory.

Table 1: Untargeted Attack Success Rates Across Different Architectures and Perturbation Budgets

Target Model Architecture	Baseline Accuracy	Success Rate (Budget Low)	Success Rate (Budget Med)	Success Rate (Budget High)
Transformer Base Variant	78.4%	34.2%	68.7%	94.1%
Transformer Large Variant	82.1%	29.8%	61.5%	89.3%
Convolutional Hybrid	76.9%	41.5%	77.2%	97.8%
Latent Diffusion Encoder	N/A	45.3%	82.1%	99.4%

We establish strict operational boundaries for the perturbation algorithms to ensure the evaluations remain practical and comparable. For iterative gradient-based attacks, we cap the maximum number of optimization steps to prevent infinite processing loops while ensuring sufficient convergence. The perturbation budgets, which dictate the maximum allowable modification to any single pixel, are strictly enforced using mathematical projection operations after every iteration. We evaluate models across three distinct perturbation tiers, designated as low, medium, and high budgets, to capture a comprehensive robustness profile. Furthermore, the generation of adversarial prompts for text-conditioned models is constrained by token length limits to ensure the malicious prompts remain syntactically coherent and realistic representations of potential real-world attacks. These implementation constraints guarantee that our evaluation framework produces reliable, scalable, and practically relevant assessments of adversarial vulnerability.

5. Results and Discussion

5.1 Vulnerability Analysis of Vision Models

The extensive empirical evaluation of vision foundation models reveals profound vulnerabilities that persist despite their massive scale and generalized pre-training. Our analysis demonstrates that these models remain highly susceptible to gradient-based adversarial perturbations, with the attack success rates scaling rapidly as the perturbation budget increases. When subjected to white-box untargeted attacks, we observe a catastrophic degradation in zero-shot classification and retrieval performance. The carefully crafted noise forces the transformer's self-attention mechanisms to focus on spurious, high-frequency artifacts rather than the semantic content of the image, effectively destroying the cross-modal alignment learned during pre-training.

Interestingly, our comparative analysis uncovers distinct robustness profiles across different model scales. While larger foundation models generally exhibit higher baseline accuracy on clean data, they do not inherently provide proportional resistance to adversarial examples. In fact, in certain targeted attack scenarios, larger models demonstrated increased vulnerability. We hypothesize that the immense parameter space of these larger models creates a denser manifold of decision boundaries, inadvertently providing adversaries with more optimized pathways to induce misclassification [32]. The increased capacity allows the model to learn highly complex, yet fragile, feature representations that are easily disrupted by precise mathematical manipulation. This finding challenges the prevailing assumption that simply scaling model size is a viable strategy for achieving adversarial robustness, indicating that fundamental architectural or training modifications are necessary to secure these systems.

5.2 Perturbation Impact on Generative Fidelity

The application of our evaluation framework to generative artificial intelligence systems exposes a complex interplay between adversarial perturbations and generative fidelity. When targeting the image conditioning inputs of latent diffusion models, we observed that even severely constrained perturbations can drastically alter the final synthesized output. In targeted attack scenarios, where the adversary aims to force the generation of a specific object, the models frequently yielded completely to the adversarial objective, producing high-fidelity images of the target concept while entirely ignoring the semantic meaning of the original, unperturbed input.

Table 2: Degradation of Generative Semantic Fidelity Under Targeted Adversarial Attacks

Generative Model Variant	Clean Image Semantic Alignment	Attacked Image Semantic Alignment	Structural Distortion Metric
Standard Diffusion Base	0.89	0.22	0.74
High-Resolution Diffusion	0.91	0.18	0.81
Accelerated Sampler Variant	0.85	0.29	0.65
Conditioned Control Model	0.94	0.15	0.88

Beyond visual manipulation, our investigation into prompt-based attacks highlights a severe vulnerability in the textual conditioning mechanisms. By appending subtly optimized adversarial token sequences to legitimate prompts, we were able to consistently bypass the integrated safety filters of the generative models. This allowed for the synthesis of restricted or inappropriate content, demonstrating a critical failure in the semantic guardrails of the system. The impact of these perturbations is not merely a reduction in image quality, but a fundamental subversion of the model's intended operational parameters. The generative models appear to overly prioritize the adversarial tokens within the cross-attention layers, allowing malicious input to dominate the reverse diffusion trajectory and dictate the final output structure.

5.3 Transferability and Defensive Implications

A critical dimension of our robustness evaluation involves assessing the transferability of adversarial examples across different foundation models. Transferability poses a severe practical threat, as it enables black-box attacks where an adversary optimizes perturbations on a widely available open-source model and successfully deploys them against proprietary, inaccessible commercial systems. Our results indicate an alarmingly high rate of transferability among models sharing similar pre-training corpora or architectural paradigms. Perturbations crafted to disrupt the cross-modal alignment of a specific contrastive learning model frequently caused significant performance degradation in entirely different transformer-based architectures. This high transferability suggests that vision foundation models, despite architectural variations, often converge upon similar latent representations and inherit shared vulnerabilities from massive, uncurated training datasets.

The profound vulnerabilities and high transferability rates observed in this evaluation have significant implications for the development of defensive mechanisms. Traditional adversarial training, which involves augmenting the training dataset with generated adversarial examples, is computationally prohibitive when applied to foundation models due to their enormous scale and the complexity of the pre-training objectives. Furthermore, reactive defenses such as input filtering or perturbation purification often degrade the model's performance on clean data and can be easily bypassed by adaptive adversaries. Our findings strongly suggest that post-hoc defenses are insufficient for securing generative artificial intelligence and vision foundation models. Instead, robustness must be integrated as a fundamental design principle during the architectural development and pre-training phases. This may involve exploring novel regularization techniques that penalize reliance on fragile high-frequency features, or developing new self-attention mechanisms that are inherently resistant to structural manipulation, ensuring the secure deployment of future artificial intelligence systems.

6. Conclusion

6.1 Summary of Findings

This paper has presented an exhaustive evaluation of the adversarial robustness of vision foundation models and generative artificial intelligence systems. Through the establishment of a rigorous methodological framework encompassing diverse threat models and advanced perturbation generation strategies, we have systematically documented the profound vulnerabilities inherent in contemporary state-of-the-art architectures. Our empirical analysis conclusively demonstrates that despite their impressive generalization capabilities and massive scale, these models remain highly susceptible to both input space modifications and semantic prompt injections. The experimental results highlight that adversarial attacks not only degrade zero-shot classification performance but also fundamentally subvert the structural integrity and semantic fidelity of generative outputs. Furthermore, the alarming rate of adversarial transferability observed between distinct models underscores the systemic nature of these vulnerabilities, suggesting that modern pre-training paradigms inadvertently cultivate shared security flaws.

6.2 Limitations and Future Directions

While this study provides critical insights into the adversarial landscape of foundation models, it is subject to certain limitations. The computational constraints associated with evaluating billion-parameter models necessitated the use of subset datasets and restricted optimization budgets, which may not capture the absolute limits of adversarial capabilities. Additionally, the rapid evolution of generative architectures means that novel attack vectors are continuously emerging, requiring ongoing methodological adaptation. Future research must prioritize the development of scalable, intrinsically robust architectures that do not rely solely on computationally prohibitive adversarial training. Investigating the theoretical bounds of robustness in cross-modal latent spaces and developing certified defensive mechanisms for diffusion processes represent crucial next steps. By continuously refining our evaluation methodologies and pursuing fundamentally secure model designs, the academic community can ensure that the unprecedented capabilities of generative artificial intelligence are deployed safely and responsibly in the future.

References

1. Liang, R., Xu, S., Xie, C., Chen, J., Ren, F., Yang, S., & Yabe, T. (2026). Abstain Mask Retain Core: Time Series Prediction by Adaptive Masking Loss with Representation Consistency. Advances in Neural Information Processing Systems, 38, 99445-99471.
2. Li, B., Zhang, R., Liang, H., Zhang, J., Zhang, J., Chen, X., ... & Wang, J. (2026). Interagent: Physics-based multi-agent command execution via diffusion on interaction graphs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 15253-15265).
3. Yao, Y., Shu, F., Li, Z., Cheng, X., & Wu, L. (2023). Secure transmission scheme based on joint radar and communication in mobile vehicular networks. IEEE transactions on intelligent transportation systems, 24 (9), 10027-10037.
4. Zhang, Y., Zhang, Y., Zhao, M., Feng, S., & Cheung, Y. M. (2026). How to Ask the AI: A User Perspective Survey for Large Language Model Prompting.IEEE Transactions on Artificial Intelligence.
5. Li, X., Wang, P., Li, G., & Zhang, Y. (2023). Design of a novel self-test-on-chip interface ASIC for capacitive accelerometers. IEEE Transactions on Circuits and Systems I: Regular Papers, 70(7), 2834-2843.
6. Wang, S., Zhang, X., Wang, P., & Li, X. (2026). Design of Magnetic Sensor Array-Based PUFs for IoT Security. IEEE Transactions on Instrumentation and Measurement, 75, 1-9.
7. Liu, C., Zhang, J., Zhang, T., Wang, Y., Zhou, H., & Jin, Q. (2026). HRIBench: Benchmarking Interaction-Centric Human-Robot Collaboration.arXiv preprint arXiv:2607.13056.
8. Jin, H., Tsai, F. S., & Martinez-Vazquez, J. (2025). Optimal expenditure decentralization for sustainable development: evidence from a 52-country panel analysis. Financial Innovation, 11(1), 111.
9. Li, X., Wang, P., Li, G., Ni, L., & Zhang, Y. (2023). Design of interface circuits and lightweight PUF for TMR sensors. IEEE Sensors Journal, 23(11), 11754-11761.
10. Zhao, Q., & Huang, Y. (2026). DAREA: A Dynamic AI-Driven Architecture for Real Estate Asset Tokenization on Blockchain.Fundamental Scientific Reports in Multidisciplinary Areas,2(02), 182-192.
11. Hu, X., Wan, Z., & Yin, Q. (2025). Token Design for Favor Trading with Dynamic Membership. Available at SSRN 6175079.
12. Issa, M. A., Chen, H., Wang, J., & Imani, M. (2024). CyberRL: Brain-inspired reinforcement learning for efficient network intrusion detection.IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,44(1), 241-250.
13. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.
14. Qu, D., Ma, Y., Yan, J., & Pyrozhenko, M. (2026). TriMeta-BFNet: A Tri-Meta Stacked Atypical-Frequency Bayesian Fourier Neural Network for Hallucination-Resistant Community Detection. Mathematics, 14 (13), 2283.
15. Ma, Y., & Qu, D. (2026, February). IMFM-GNN: Interpretable Multimodel GNN Via Fourier-Markov Dynamics for Community Detection in Sustainable Energy Networks. In 2026 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA) (pp. 1-6). IEEE.
16. Wuyang Zhang and Shichao Pei. 2026. Your LLM Agent Can Leak Your Data: Data Exfiltration via Backdoored Tool Use. arXiv:2604.05432 [cs.CR] doi:10.48550/arXiv.2604.05432

17. Yang, W., Qin, Y., & Yang, Y. (2021). Characterizing the multicast nature of malicious flows in CCN via SIS epidemic model. IEEE Systems Journal, 16(2), 2240-2250.
18. Yang, P., Chen, W., Wang, K., Ai, L., Yang, E., & Shi, T. (2026, July). EVM-QuestBench: An Execution-Grounded Benchmark for Natural-Language Transaction Code Generation. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 35513-35529).
19. Tao, J., Lyu, R., & Cao, X. (2026). A Deep Learning-Based Automated Content Moderation Framework for Online Platforms.Future-Adaptive Intelligence and Lifelong Systems,1(1).
20. Luo, P., & Ding, X. (2026). C-MAPPO-TSC: A Collaborative Multi-Agent Reinforcement Learning Framework for Urban Traffic Signal Control and Congestion Mitigation. Fundamental Scientific Reports in Multidisciplinary Areas, 2 (02), 170-181.
21. Hu, X., Wan, Z., & Murthy, N. N. (2019). Dynamic pricing of limited inventories with product returns. Manufacturing & Service Operations Management, 21(3), 501-518.
22. Yang, Z., Ji, W., & Wang, Z. (2024). Adaptive joint configuration optimization for collaborative inference in edge-cloud systems. Science China Information Sciences, 67(4), 149103.
23. Long, J., Xu, Z., Jiang, T., Yao, W., Jia, S., Ma, C., & Chen, X. (2025, April). Robust SAM: on the adversarial robustness of vision foundation models. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 6, pp. 5775-5783).
24. Yang, Z., Liang, B., & Ji, W. (2021). An intelligent end–edge–cloud architecture for visual IoT-assisted healthcare systems. IEEE internet of things journal, 8(23), 16779-16786.
25. Zhang, Q., Gao, W., Liu, C., Yao, Y., Yan, S., Shu, F., & Jin, S. (2025). Covert transmission for active RIS-aided full-duplex UAV integrated sensing, communication, and computation systems. IEEE Journal on Selected Areas in Communications.
26. Yao, Y., Zhao, J., Li, Z., Cheng, X., & Wu, L. (2023). Jamming and eavesdropping defense scheme based on deep reinforcement learning in autonomous vehicle networks. IEEE Transactions on Information Forensics and Security, 18, 1211-1224.
27. Yang, Z., Ji, W., Guo, Q., & Wang, Z. (2023, October). Javp: Joint-aware video processing with edge-cloud collaboration for dnn inference. In Proceedings of the 31st ACM International Conference on Multimedia (pp. 9152-9160).
28. Qu, D., Ma, Y., & Wang, Y. (2026). STR-GNN: stability-regularized graph neural networks for suppressing spurious temporal variations in dynamic community detection. Scientific Reports.
29. Liao, Qiuyue, et al. "Explainable Artificial Intelligence for 5G Security and Privacy: Trust, Governance, and Resilience." (2025).
30. Ma, Y., Qu, D., & Wang, Y. (2026). HALO-GNN: hallucination-resistant temporal graph neural networks for dynamic community detection.Scientific Reports.
31. Tao, J., Lyu, R., & Cao, X. (2026). A Scalable Data Governance Architecture for Privacy-Aware Intelligent Learning Systems in Lifelong. Future-Adaptive Intelligence and Lifelong Systems, 1 (1).
32. Zhang, H., Zandsalimy, M., & Sushmita, S. (2026). Exposing LLM Safety Gaps Through Mathematical Encoding: New Attacks and Systematic Analysis.arXiv preprint arXiv:2605.03441.