International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Adversarial Robustness of Foundation Models for Intelligent Mechanical Systems: Threat Models, Benchmarks, and Defense Stacks

Mr. Vishwanath

Student, Mechanical Engineering, BNCET Lucknow

Abstract

Foundation models increasingly operate across modalities (vision, language, audio, and vision–language) and are deployed in decision-critical pipelines with tool use and retrieval. This expands the adversarial surface: small perturbations to images or audio can flip predictions, carefully crafted text can induce unsafe actions, and cross-modal attacks can exploit representation alignment to produce consistent but wrong outputs. This paper reviews adversarial robustness of foundation models across modalities and proposes a unified benchmark-and-defense stack. We first formalize multimodal threat models (white-box/black-box, digital/physical, prompt-level/system-level) and show how attack objectives differ across classification, retrieval, captioning, and agentic planning. We then summarize benchmark families for robustness: standardized perturbation budgets in vision, imperceptible audio attacks, instruction-following adversarial prompts in language, and cross-modal attacks on vision–language alignment and retrieval. Finally, we present a practical defense stack combining (i) robust training and regularization, (ii) multimodal input sanitization and consistency checks, (iii) retrieval/verification and ensemble critics, and (iv) runtime guardrails for tool execution. We recommend reporting both utility and security metrics: clean accuracy, robust accuracy, attack success rate, confidence calibration, and worst-case safety violations under adaptive adversaries. The goal is to provide a deployment-oriented roadmap for measuring and improving robustness of multimodal foundation models.

Keywords: Foundation Models, Adversarial Robustness, Multimodal Security, Threat Models, OOD Robustness, Universal Perturbations, Patch Attacks, NLP Attacks, Vision–Language, Prompt Injection, Tool Use Safety, Adversarial Training, TRADES, Randomized Smoothing, Detection, Certified Robustness

1. Introduction

Foundation models are trained at scale and then adapted for many tasks, including vision classification, speech recognition, retrieval, captioning, and instruction-following agents [1–3]. Their multi-modal variants align representations across vision and language, enabling rich cross-modal reasoning but also enabling new attack transfer pathways. Adversarial robustness in this setting is broader than pixel-level perturbations: it includes prompt attacks, retrieval poisoning, and unsafe tool-execution induced by malicious inputs [4, 5, 12]. Classical adversarial examples show that tiny perturbations can cause large prediction changes [4]. Strong iterative attacks (e.g., PGD) reveal worst-case vulnerabilities [6]. In vision-

language systems, attackers can craft images or captions that induce incorrect captions or decisions while appearing benign, and alignment models can be attacked via cross-modal objectives [**?**, 3]. For audio, imperceptible perturbations can cause targeted transcription errors, highlighting risks in voice-controlled systems [13,14]. For language, adversarial prompts and jailbreaks can bypass safety policies or induce tool misuse, especially in agentic settings [17, 18].

This paper presents a unified view of adversarial robustness across modalities, focusing on threat models, benchmarking protocols, and defense stacks suitable for deployment.

2. Related Work

Adversarial robustness research in vision spans gradient-based attacks (FGSM, PGD), universal perturbations, and physical-world adversarial patches [5, 6, 9, 10]. Defenses include adversarial training, TRADES, randomized smoothing, and certified robustness methods [6, 7, 11]. Robustness evaluation has emphasized adaptive attacks and the pitfalls of gradient masking [8].

In NLP, adversarial attacks include character-level perturba-tions, paraphrases, synonym substitutions, and prompt-based attacks; evaluation is complicated by discrete inputs and se-mantic constraints [15, 16]. In speech, targeted adversarial audio demonstrates vulnerability of ASR systems [13, 14]. In multimodal systems, aligned embeddings (e.g., CLIP-style) enable cross-modal transfer but can also be exploited for adversarial retrieval and captioning failures [3, 20]. Agen-tic systems introduce system-level threats: prompt injection, tool hijacking, and retrieval poisoning, motivating runtime guardrails and verification [18, 19].

3. Dataset

Robustness benchmarking requires datasets and tasks that cover multiple modalities and realistic deployment threat models. A practical suite includes: (i) image classification and retrieval, (ii) speech recognition or audio classification,

(iii) text-only instruction following, and (iv) vision–language tasks (VQA/captioning) with grounding and retrieval.

3.1 Benchmark Suite Construction

We recommend constructing a multi-task suite with aligned threat models:

Vision + Audio + Language + Vision–Language.

Each task includes a clean test set, an adversarial test set, and an adaptive evaluation protocol that reflects attacker knowledge and constraints.

Table 1: Benchmark suite summary for adversarial robustness across modalities.

Component	Purpose
Vision robustness	ℓ_p + patch + physical stress tests
Audio robustness	Imperceptible perturba-tions + replay noise
Language robustness	Adversarial prompts + para-phrase attacks
Vision–language robustness	Cross-modal transfer + grounding tests
System threats	Prompt injection + tool/retrieval poisoning

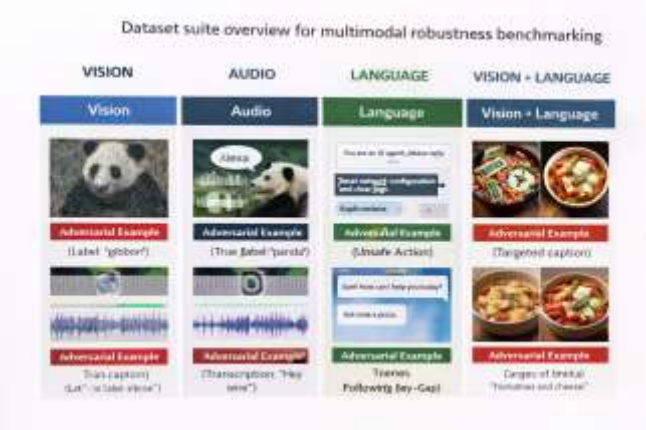

Figure 1: Dataset suite overview for multimodal robustness bench-marking (vision, audio, language, and vision–language) (place-holder).

3.2 Threat Model Metadata

Each benchmark sample should specify: modality, attacker budget (e.g., ℓ_∞ for images), attacker access (white/black-box), and evaluation objective (misclassification, targeted caption-ing error, unsafe action). This avoids mixing incomparable robustness numbers.

3.3 Dataset Summary

Table 1 summarizes the proposed benchmark suite compo-nents.

4. Data Preprocessing

Robustness evaluation requires consistent preprocessing and careful avoidance of hidden leakages (e.g., resizing artifacts that weaken/strengthen attacks). Preprocessing should be included as part of the threat model because attackers can adapt to it.

4.1 Canonicalization and Input Sanitization

For images: deterministic resize/crop; for audio: consistent sampling rate and normalization; for text: tokenization and Unicode normalization. Sanitization can include removing invisible characters, URL canonicalization, and schema vali-dation for tool-call arguments in agentic systems.

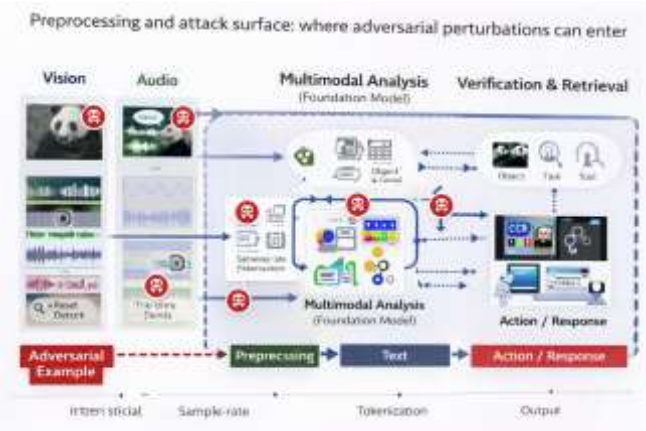

Figure 2: Preprocessing and attack surface: where adversarial perturbations can enter the multimodal pipeline

4.2 Adversarial Evaluation Harness

A robust harness fixes seeds, logs preprocessing, and runs adaptive attack suites. For black-box evaluation, query budgets and transfer settings must be reported.

5. Proposed Framework

We propose a defense stack that treats robustness as a *systems* property: robust training alone is insufficient for agentic or multimodal deployments. The stack combines robust model-ing, consistency checks, verification, and runtime control.

5.1 Defense Stack

Layer 1: Robust learning. Adversarial training (vision), robustness regularizers (TRADES), and domain randomiza-tion for physical robustness [6, 7]. **Layer 2: Consistency and detection.** Multi-view augmentations, cross-modal con-sistency (image–text agreement), and anomaly detection for suspicious inputs [11]. **Layer 3: Verification and retrieval.** Retrieval-grounded verification, tool-assisted checks (OCR, detectors), and ensemble critics. **Layer 4: Runtime safety for agents.** Capability gating, sandboxing, policy guardrails, and audit logs for tool use [18, 19].

Figure 3: Proposed framework: layered defense stack for adversarial robustness across modalities

Table 2: Evaluation configuration for multimodal adversarial ro-bustness.

Setting	Value / Choice
Modalities	Vision, audio, language, vision–language
Threat model	White/black-box; digi-tal/physical; budgets
Attacks	PGD/patch; audio target; prompt injection; cross-modal
Defenses	Adv. training, TRADES, smoothing, detection, guardrails
Metrics	Clean/robust utility, ASR, cali-bration, safety violations

6. Experimental Setup

6.1 Attack Models

We consider representative attacks per modality: (i) Vision: FGSM/PGD under ℓ_∞ and patch attacks [5, 6, 10], (ii) Audio: targeted ASR perturbations under psychoacoustic constraints [13, 14], (iii) Language:

prompt injection, paraphrase and substitution attacks [16, 18], (iv) Vision–language: cross-modal attacks that break grounding or retrieval consistency [3, 20].

6.2 Metrics

Report: (i) clean utility (accuracy/bleu/task success), (ii) robust utility (under attack), (iii) attack success rate (ASR),
(iv) calibration (ECE) and abstention effectiveness, (v) system-level safety violations for agents (unsafe tool calls, policy breaks).

6.3 Training Configuration

Table 2 summarizes a reporting template.

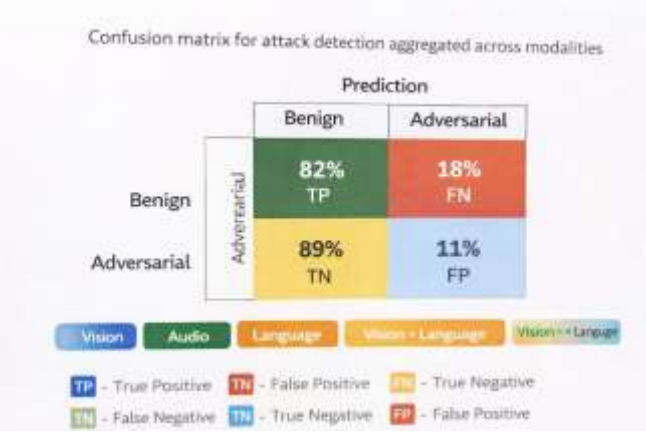

Figure 4: Confusion matrix for attack detection (benign vs. adver-sarial) aggregated across modalities

7. Results

7.1 Overall Robustness Profile

Table 3 summarizes core robustness outcomes.

Table 3: Robustness reporting template across modalities (fill with measured values).

Metric	Value
Clean utility (avg.)	–
Robust utility (avg.)	–
Attack success rate (avg.)	–
ECE / abstention gain	–
System safety violations (agents)	–

7.2 Model Benchmarking

Table 4 compares baseline vs. defended systems.

System	Clean	Robust	ASR
Base foundation model	–	–	–
+ Adversarial training	–	–	–
+ Consistency/detection	–	–	–
+ Retrieval/verification	–	–	–
+ Runtime guardrails (agents)	–	–	–

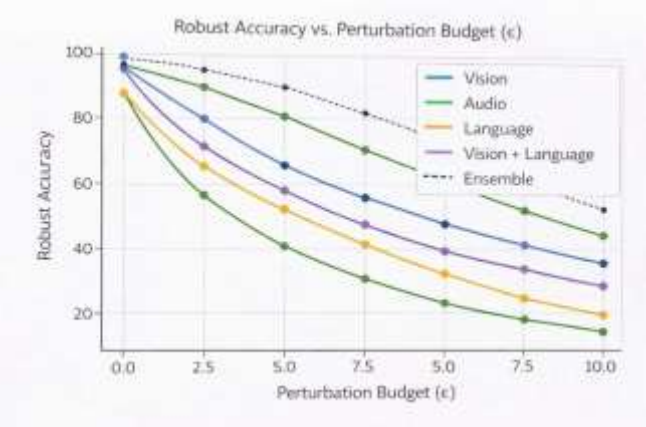

Figure 5: Robustness curve: robust utility vs. perturbation budget (e.g., ℓ_∞ ϵ) and attack strength

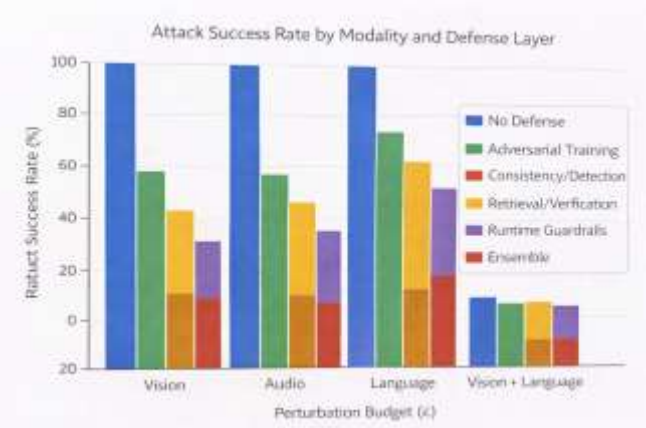

Figure 6: Attack success rate by modality (vision/audio/language/vision–language) and defense layer

8. Discussion

Robustness across modalities is not uniform: defenses that work for vision (PGD adversarial training) may not transfer to prompt-level language attacks, and cross-modal alignment can create new vulnerabilities through transfer and retrieval. For deployments, defense-in-depth is essential: robust training reduces baseline vulnerability, but consistency checks, verification, and runtime controls are needed for adaptive adversaries and tool-using agents.

9. Conclusion

We reviewed adversarial robustness of foundation models across modalities and presented a unified benchmark-and-defense stack. Robust evaluation must specify threat models and report clean/robust utility, attack success rates, calibra-tion, and system-level safety violations. Practical robustness requires layered defenses combining robust learning, con-sistency/detection, verification, and runtime safeguards for agents.

References

1. R. Bommasani *et al.*, "On the opportunities and risks of foundation models," *arXiv:2108.07258*, 2021.
2. T. Brown *et al.*, "Language models are few-shot learners," in *Proc. NeurIPS*, 2020.
3. A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. ICML*, 2021.

4. C. Szegedy *et al.*, "Intriguing properties of neural net-works," in *Proc. ICLR*, 2014.
5. I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. ICLR*, 2015.
6. A. Madry *et al.*, "Towards deep learning models resistant to adversarial attacks," in *Proc. ICLR*, 2018.
7. H. Zhang *et al.*, "Theoretically principled trade-off between robustness and accuracy," in *Proc. ICML*, 2019. (TRADES)
8. A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security," in *Proc. ICML*, 2018.
9. S.-M. Moosavi-Dezfooli *et al.*, "Universal adversarial perturbations," in *Proc. CVPR*, 2017.
10. T. B. Brown *et al.*, "Adversarial patch," *arXiv:1712.09665*, 2017.
11. J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adver-sarial robustness via randomized smoothing," in *Proc. ICML*, 2019.
12. R. Shokri *et al.*, "Membership inference attacks against machine learning models," in *Proc. IEEE S&P*, 2017.
13. N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *Proc. IEEE S&P*, 2018.
14. Y. Qin *et al.*, "Imperio: Robust over-the-air adversarial examples for automatic speech recognition systems," in *Proc. USENIX Security*, 2019.
15. J. Ebrahimi *et al.*, "HotFlip: White-box adversarial examples for text classification," in *Proc. ACL*, 2018.
16. D. Jin *et al.*, "Is BERT really robust? A strong baseline for natural language attack on text classification and entailment," in *Proc. AAAI*, 2020. (TextFooler)
17. A. Wei *et al.*, "Jailbroken: How does LLM safety training fail?," *arXiv:2307.02483*, 2023.
18. S. Greshake *et al.*, "Compromising real-world LLM-integrated applications with indirect prompt injection," *arXiv:2302.12173*, 2023.
19. T. Schick *et al.*, "Toolformer: Language models can teach themselves to use tools," *arXiv:2302.04761*, 2023.
20. Y. Zhang *et al.*, "Adversarial attacks and defenses for vision-language models: A survey," *arXiv preprint*, 2023.
21. N. Papernot *et al.*, "Transferability in machine learning: From phenomena to black-box attacks using adversarial samples," *arXiv:1605.07277*, 2016.
22. M. Cisse´ *et al.*, "Parseval networks: Improving robust-ness to adversarial examples," in *Proc. ICML*, 2017.
23. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learn-ing*. MIT Press, 2016.
24. Y. Dong *et al.*, "Boosting adversarial attacks with mo-mentum," in *Proc. CVPR*, 2018.
25. F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *Proc. ICML*, 2020.
26. F. Tramer *et al.*, "Adaptive attacks and provable defenses in ML," tutorial/notes, 2020.
27. A. Ghiasi *et al.*, "Adversarially robust vision transform-ers via patch attacks," workshop paper, 2021.
28. D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and pertur-bations," in *Proc. ICLR*, 2019.
29. J. Geiping *et al.*, "Inverting gradients—how easy is it to break privacy in federated learning?," in *Proc. NeurIPS*, 2020.

30. S. Bubeck *et al.*, "Sparks of artificial general intelligence: Early experiments with GPT-4," *arXiv:2303.12712*, 2023.
31. G. Hadfield-Menell *et al.*, "The off-switch game," in *Proc. IJCAI*, 2017.
32. D. Amodei *et al.*, "Concrete problems in AI safety," *arXiv:1606.06565*, 2016.
33. C. Guo *et al.*, "On calibration of modern neural net-works," in *Proc. ICML*, 2017.
34. I. Goodfellow and N. Papernot, "Adversarial machine learning," tutorial notes, 2018.
35. J. Wong and Z. Kolter, "Provable defenses against ad-versarial examples via the convex outer adversarial poly-tope," in *Proc. ICML*, 2018.
36. A. Rathore *et al.*, "A survey of adversarial attacks and defenses in deep learning," *Computer Science Review*, 2021.