Review 1 on
Explainable AI for Robust Defense Against Adversarial Image Attacks

In Partial fulfilment for the Requirements of Project on Machine Learning

By
Mohammed Ashlab 23MIA081
Rayini Amarendar Reddy 23MIA1012
Integrated Mtech CSE with Business Analytics

Under the guidance of
Dr.Abdul Quadir Md SCOPE

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

September 2025

Explainable AI for Robust Defense Against Adversarial Image Attacks

Abstract

This research focuses on developing a severity-aware, lightweight defense mechanism against prompt and multimodal injection attacks in Large Language Models (LLMs) and Vision-Language Models (VLMs). While existing approaches achieve high detection accuracy, they are often computationally heavy and unsuitable for real-time applications. Our framework integrates severity classification, adaptive defense strategies, and Explainable AI (XAI) components to balance security, efficiency, and transparency. This document reviews existing literature, identifies key research gaps, formulates the problem statement, and defines research objectives.

Introduction

Prompt and multimodal injection attacks pose significant security threats to modern AI systems. Attackers can embed malicious instructions in text, images, or cross-modal inputs to manipulate model outputs. While recent research proposes defenses such as adversarial training, data sanitization, and robust XAI, these approaches often lack adaptability, impose high computational costs, and fail to provide transparent justifications for defensive actions. This research proposes a severity-aware defense mechanism that adapts responses based on input severity and explains decisions in a user-friendly way. The goal is to enable practical, lightweight, and trustworthy defenses for real-world deployments.

1.Literature Survey

S.No	Paper title	Year	Summary	Methodology	Pros	Cons	ML concept used	DL concept used
1	BadCLIP: Dual-Embedding Guided Backdoor Attack on Multimodal Contrastive Learning	2024	Backdoor triggers flip CLIP alignment and bypass several detection and unlearning defenses.	Dual-embedding trigger optimization targeting contrastive objectives in CLIP encoders.	Strong transfer and stealth on CLIP-like encoders.	Focus on contrastive VL pretraining, not generative LVLMs.	Backdoor benchmarking, defense bypass analysis.	Contrastive VL training, latent trigger optimization.
2	Backdooring Multimodal Learning	2024	Systematizes multimodal backdoor vulnerabilities from fusion and modality weighting.	Analytical threat modeling across fusion stages and heterogeneous modalities.	Security and privacy framing guides defense mapping.	Conceptual: requires engineering for practice.	Defense surface mapping and audit checklists.	Multimodal fusion architectures and modality patterns.
3	Benchmarking and Defending	2025	Introduces systematic benchmarks	Agent evaluation with hidden-state	Agent-centric, realistic	LLM-centric with limited multimodal scope.	Red teaming protocols,	Hidden-state classifiers and

	Against Indirect Prompt Injection in LLM Agents		and defenses for agents vulnerable to indirect prompt injection.	detectors and prompt shields.	evaluation protocols.		retrieval/tool risk modeling.	middleware filtering.
4	IDEATOR: Jailbreaking and Benchmarking Large Vision-Language Models Using Themselves	2024	Automated multimodal jailbreak generation with strong black-box transfer to LVLMs.	Red teaming with diffusion-generated adversarial contents.	Multimodal jailbreak focus with transferability.	Preprint status with evolving benchmarks.	Benchmark creation and ASR metric design.	VLM-driven adversarial attack generation.
5	JailBreakV: A Benchmark for Assessing the Robustness of Multimodal LLMs	2024	Large-scale jailbreak transferability and robustness benchmark.	Curated safety tasks for comparative evaluation.	Broad comparisons.	Evolving metrics and curation.	Robustness scoring and comparative analysis.	Multimodal alignment stress tests.
6	Adaptive Prompt Injection Challenge (LLMail-Inject) — Competition Report	2025	Realistic agent scenarios revealing defense gaps.	Multi-defense evaluation with level-based scenarios.	Production-like insights.	Competition report vs formal paper.	Attack adaptation analysis, benchmarking.	Ensemble and hidden-state classifiers.
7	Lessons from Defending Gemini Against Indirect Prompt Injection — Whitepaper	2025	Field-tested layered mitigations for indirect prompt injection.	Defense-in-depth patterns and incident response loops.	Practical deployment insights.	Not peer-reviewed.	Severity labeling and incident response playbooks.	Instruction filtering, tool-call mediation etc.
8	Explainable AI in Medical Imaging: Beyond Saliency-Based Approaches	2023	Surveys concept attribution and counterfactuals for deeper interpretability.	Comparative taxonomy and evaluation principles.	Rich XAI toolbox.	Added complexity for validation/deployment.	Explanation evaluation frameworks and user studies	Concept bottlenecks, prototypes, counterfactuals.
9	A Systematic Review of Explainable AI in Medical Image Analysis	2024	Comprehensive survey on imaging XAI methods and trends.	Systematic screening and taxonomy building.	Broad coverage and deployment insights.	Heterogeneous metrics in studies.	Deployment constraints and evaluation frameworks.	Intrinsic/post-hoc explainers, attention mechanisms.
10	Explainable Artificial Intelligence for Medical Applications	2024	Multimodal XAI survey linking transparency with safety and auditability.	Cross-modality synthesis with taxonomy and evaluation recommendations.	Broad applicability.	High-level synthesis, limited detail.	Safety-by-design and audit protocols.	Saliency, attention, concept/prototype methods.
11	Robust Evaluation of Diffusion-Based Adversarial Purification	2023	Highlights evaluation pitfalls and robust protocols for	Empirical stress-testing with standardized metrics.	Standardized evaluation.	Focus on vision encoders.	Robustness metric design and test-time protocols.	Diffusion purification with ViT/CNN backbones.

			diffusion purification.					
12	Enhancing Adversarial Robustness via Score-Based Generative Models	2023	Score-based diffusion guidance improves purification & certified robustness.	Reverse SDE denoising guided by score matching.	Theory-backed improved robustness.	High compute, latency overhead.	Certified robustness and tuning.	Score-based diffusion and guided denoising.
13	Adversarial Purification with One-Step Guided Diffusion	2025	Low-latency one-step guided diffusion for competitive robustness.	Single-step denoising with calibrated guidance.	Suitable for real-time systems.	Sensitive guidance calibration required.	Latency-robustness tradeoff profiling.	Latent diffusion, single-step defense.
14	Gradient-Free Adversarial Purification with Diffusion Models	2025	Gradient-free denoising for robust purification under strong attacks.	Gradient-free denoising schedule and evaluation.	Efficiency gains.	Preprint with ongoing validations.	Robust accuracy reporting.	Diffusion denoising without gradients.
15	Adversarial Attacks and Defenses in Machine Learning: A Survey	2023	System-level taxonomy of attacks and defenses for ML pipelines.	Wide synthesis of attack, defense, and evaluation strategies.	Authoritative and comprehensive.	Not specific to multimodal VLMs.	Threat modeling and evaluation standards.	Robust training, detection, purification strategies.
16	Adversarial Examples: A Survey in Deep Learning	2024	Practical cross-domain survey including vision defense insights.	Mapping attack/defense types and application constraints.	Application-focused across domains.	Mixed domain scope limits specialization.	Transferability and defense taxonomies.	Gradient/black-box attacks and robust optimization.
17	Vision-Language Models for Vision Tasks: A Survey	2024	Authoritative review mapping defense insertion points across VLM tasks.	Taxonomy across tasks, architectures, pretraining, fusion.	Deep task and architecture coverage.	Limited to few security concerns.	Task/benchmark mapping.	Contrastive/generative VL pretraining, fusion, adapters.
18	VLP: A Survey on Vision-Language Pre-training	2023	Surveys VL pretraining objectives and families.	Comparative review of CLIP, masked modeling, generative heads.	Strong Springer venue coverage.	Less security focus.	Pretraining regimes and transfer strategies.	CLIP contrastive, MLM/VQA heads, cross-modal encoders.
19	A Survey on Efficient Vision-Language Models	2025	Reviews methods for efficient on-device VLMs like quantization and pruning.	Systematic synthesis of compression and optimization techniques.	Efficient deployments.	Preprint with evolving best practices.	Latency-accuracy trade-offs and memory profiling.	Quantization, sparsity, lightweight adapters.
20	A Comprehensive Survey of Vision-Language Models	2025	Broad survey of VLM architectures and benchmarks to inform deployment.	Aggregation of model families and evaluation suites.	Up-to-date comprehensive coverage.	Paywalled details limit access.	Benchmark synthesis and trend analysis.	Multimodal encoders, fusion variants.
21	Multimodal Prompt Injection	2025	Survey of prompt injection	Empirical evaluation of attack vectors	Comprehensive coverage.	Defensive techniques still in infancy.	Adversarial attacks, security.	Vision-language models.

	Attacks: Risks and Defenses		vulnerabilities and defenses.	and defensive strategies.				
22	Manipulating Multimodal Agents via Cross-Modal Prompt Injection	2023	Cross-modal adversarial input optimization for manipulating agents.	Optimization framework generating attacks across visual and textual modes.	Novel cross-modal frameworks with black-box success.	High computational complexity.	Adversarial ML.	Vision-language models.
23	Enhancing Prompt Injection Attacks to LLMs via Poisoning	2024	Combined poisoning and prompt injection attacks improving success.	Poisoning plus prompt injection to boost attack effectiveness.	Novel combined attack framework.	Limited in unimodal text-only models.	Poisoning attacks.	LLMs.
24	LLMs for Explainable AI: A Comprehensive Survey	2020	Survey of explainability techniques for large language models.	Review of LLM-focused explainability methods.	Extensive literature coverage.	Mainly textual LLM focus, limited experiments.	Large language models.	Explainability.
25	OCR Post-Correction for Detecting Adversarial Text Images	2022	Improves OCR robustness against adversarial textual samples.	Denoising autoencoders and post-processing for OCR correction.	Significantly improves OCR accuracy under attack.	Limited to adversarial textual perturbations.	OCR robustness.	CNN denoising.

2. Research Gap

- To design and develop a **lightweight multimodal defense framework** for consumer-facing LLM and VLM assistants that can process both text and image inputs.

- To implement **intent classification and anomaly detection models** (including perturbation analysis, spectral checks, and Moondream-based image understanding) to identify malicious or hidden instructions even Multiple languages.

- To integrate **Explainable AI (XAI)** mechanisms to provide **transparent, user-friendly, and auditable explanations** for flagged content.

- To evaluate the framework's **performance in terms of detection accuracy, computational efficiency, real-time usability, and explainability**, comparing it with existing defenses against malicious prompt injections.

3. Problem Statement

Despite the growing adoption of LLM and VLM-based consumer AI assistants, most existing systems are vulnerable to hidden malicious instructions embedded in both text and images.

Current defenses are either computationally heavy or opaque, limiting their practicality for real-time applications and reducing user trust. Furthermore, the rise of AI agents with retrieval-augmented generation (RAG) capabilities introduces additional attack surfaces, as models now integrate external knowledge sources that could be exploited. Therefore, there is a critical need for a lightweight, explainable, and robust defense framework that can detect both text- and image-based malicious injections—including noise-encoded or steganographic payloads—while providing transparent explanations to users and developers in real time.

4. Research Objectives

1. To design a lightweight severity detection mechanism that classifies inputs into severity levels.
2. To implement adaptive defense strategies mapped to severity categories.
3. To integrate Explainable AI components that generate concise justifications of defense decisions.
4. To evaluate computational efficiency, accuracy, and latency against heavyweight baselines.
5. To demonstrate multimodal capability by testing on text + image attack scenarios.
6. To provide a deployment blueprint for integration into real-world LLM/VLM pipelines.

5. Conclusion

This work proposes leveraging explainable AI techniques to enhance the security and interpretability of vision–language models facing adversarial image attacks. By integrating XAI-driven analysis, the approach aims to uncover and defend against malicious manipulations while providing transparent insights into model decisions. The concept bridges interpretable AI and robust adversarial defense, offering a novel direction for developing trustworthy, secure machine learning systems in complex visual environments