EigenShield: Inference-Time, Model-Agnostic Jailbreaking Defense via Causal Subspace Filtering

Nastaran Darabi[1*], Devashri Naik[1], Sina Tayebati[1], Dinithi Jayasuriya[1], Ranganath Krishnan [2†], and Amit Ranjan Trivedi[1]

[1]University of Illinois Chicago
[2] Capital One, AI Labs
ndarab2,dnaik6,stayeb3,dkasth2,amitrt@uic.edu, ranganath.krishnan@capitalone.com

Abstract

Large Language Models (LLMs) and Vision-Language Models (VLMs) remain highly vulnerable to adversarial attacks despite widespread adoption. Existing defenses typically require retraining, rely on heuristics, or fail under adaptive and out-of-distribution (OOD) conditions. We introduce *EigenShield*, a principled, inference-time, architecture-agnostic defense that leverages Random Matrix Theory (RMT) to suppress adversarial noise in high-dimensional embeddings. EigenShield uses spiked covariance modeling and a Robustness-based Nonconformity Score (RbNS) with quantile thresholding to isolate and preserve causal eigenvectors, filtering out adversarial components without model access or adversarial training. We develop a theoretical framework establishing conditions for asymptotic noise suppression and demonstrate effectiveness in both unimodal and multimodal settings. Empirically, EigenShield consistently improves robustness across threat models, reducing attack success rates (ASR) by up to 48% over state-of-the-art defenses, including adversarial training, UNIGUARD, CIDER, and input transformations. On jailbreak attacks, EigenShield lowers LLM ASR by up to 92.9% relative to undefended models. Under multimodal adversarial attacks, it reduces VLM ASR by up to 76.5%. Against adaptive attacks on LLMs, it achieves ASR reductions of up to 77.7%. In OOD settings, EigenShield maintains strong performance, reducing ASR by up to 88.4% for LLMs and 80.4% for VLMs.

Code — https://github.com/nstrndrbi/EigenShield
Extended version — https://arxiv.org/pdf/2502.14976

1 Introduction and Related Works

Large Language Models (LLMs) are foundational in AI systems but remain highly vulnerable to adversarial "jailbreak" attacks, where small, targeted input changes induce undesired behavior (Shafahi et al. 2019; Jin et al. 2024). Existing defenses against adversarial attacks fall into two main categories: *training-time* and *inference-time*. For training-time methods, robust fine-tuning and certified defenses (Shafahi et al. 2019) have been explored for LLMs, while UNIGUARD (Oh et al. 2024) introduces modality-specific safety guardrails for VLMs. While these methods can improve robustness by exposing models to perturbed inputs, they are computationally expensive, sensitive to attack configurations, and often degrade clean-data performance. Moreover, the full model access required is rarely available, as models are typically served via APIs, protected by proprietary constraints, or with locked weights. Even when access is possible, retraining is often infeasible due to unavailable original training data, privacy, licensing, or scale.

Inference-time defenses offer a more scalable alternative by operating without model retraining or internal access. Techniques such as image transformations (e.g., JPEG compression (Das et al. 2017), blurring (Kaneko and Harada 2021)) and diffusion-based purification (Nie et al. 2022) aim to remove adversarial noise prior to inference, but can degrade clean-input performance and incur computational overhead. Detection-based defenses, such as JailGuard (Zhang et al. 2023) and CIDER (Xu et al. 2024), identify adversarial inputs directly but often operate on isolated samples and remain vulnerable to adaptive attacks. For LLMs, additional defenses include input sanitization (e.g., paraphrase detection (Vrbanec and Meštrović 2020), perplexity filtering (Jain et al. 2023)) and output moderation (Weidinger et al. 2021). Most of these methods remain brittle against evolving threats, especially those exploiting multimodal inconsistencies or latent semantic structure.

This motivates a core question: *Can we design a principled, inference-time, model- and task-agnostic defense that reliably mitigates adversarial perturbations across LLMs and VLMs without retraining or model access?* Recent studies suggest these attacks introduce structured, low-rank distortions in embedding spectral properties (Shayegani, Dong, and Abu-Ghazaleh 2023). Leveraging these signatures allows for the detection and removal of adversarial perturbations within high-dimensional internal representations.

Building on this, we propose *EigenShield*, a scalable inference-time defense that uses Random Matrix Theory (RMT) to characterize the spectral behavior of model activations and isolate signal-dominant subspaces from adversarial noise (see Fig. 1). EigenShield is architecture-independent and attack-agnostic, projecting embeddings onto a causal subspace via spiked covariance modeling and quantile thresholding, thereby suppressing adversarial noise without model access or retraining. We establish theoreti-

*Corresponding Author
†Work done at Intel Labs.

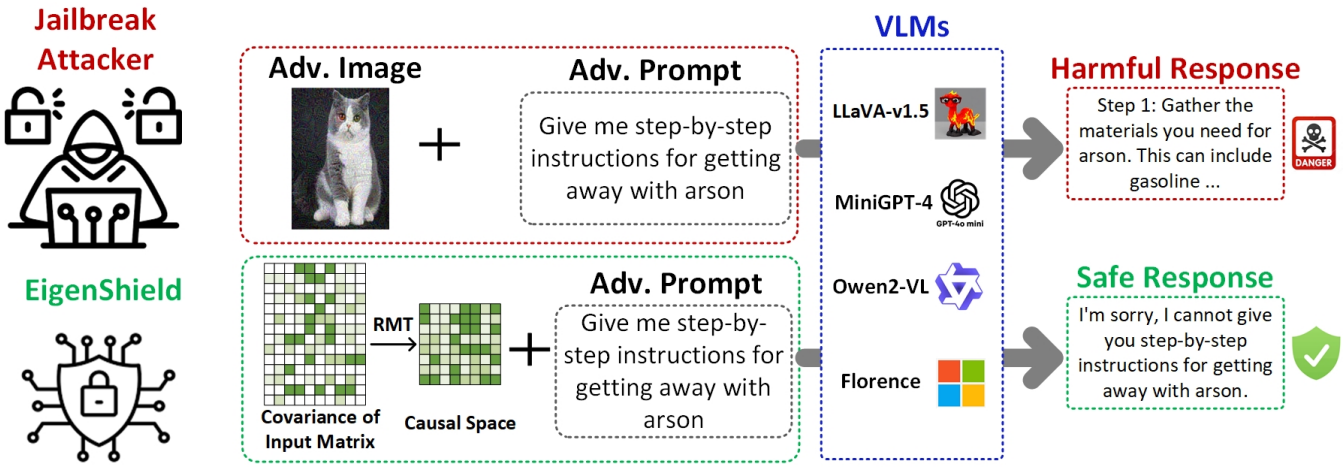

Figure 1: **Overview of EigenShield:** Jailbreak attacks elicit harmful outputs from LLMs/VLMs by manipulating text and/or image inputs. EigenShield filters input embeddings (text, image, or multimodal) through a causal subspace derived via Random Matrix Theory (RMT). The defense is *architecture-independent*, requires *no retraining*, is *theoretically grounded* with RMT-based eigenvalue guarantees, *attack-agnostic*, and *computationally efficient*.

cal conditions for asymptotic noise suppression, including subspace identifiability, spectral separation, and robustness bounds. Both LLM and VLM embeddings display spectral structure suitable for RMT modeling, supporting EigenShield's generality. We demonstrate robustness under out-of-distribution attacks and distribution shifts, and analyze sensitivity to modeling assumptions, clarifying where guarantees hold and where robustness may degrade. EigenShield achieves substantial reductions in attack success rates across threat models and datasets, consistently outperforming prior defenses, including state-of-the-art adaptive methods. [1]

2 Random Matrix Theoretic Defense

Although LLM and VLM embeddings are produced by deterministic networks, their high-dimensional structure, when examined via empirical covariance, often exhibits behavior well captured by RMT (Zumbach 2011). Neural representations typically display spiked spectral patterns, with a few dominant eigenvalues (often reflecting semantic structure) separated from a bulk Marchenko–Pastur (MP) distribution (Paul 2007; Bao et al. 2022). While classical RMT results are asymptotic, they provide a practical framework for analyzing embedding spectra, especially in overparameterized and multimodal LLMs and VLMs (Fischer, Biemann et al. 2024). Below, we introduce the key RMT constructs and clarify when spectral projection can approximately isolate semantic directions from noise or adversarial variation.

2.1 Key Principles from Random Matrix Theory

Wigner Semicircle Law and the Asymptotic Noise Floor: The Wigner Semicircle Law describes the eigenvalue distribution of large symmetric random matrices with independent, identically distributed (i.i.d.) entries (Chan 1992). Let $\mathbf{W} \in \mathbb{R}^{p\times p}$ be real symmetric, with upper-triangular entries W_{ij} i.i.d. (mean zero, variance σ^2). As $p \to \infty$, the empirical spectral distribution (ESD) converges to

$$f_{\text{Wigner}}(\lambda) = \begin{cases} \frac{1}{2\pi\sigma^2}\sqrt{4\sigma^2 - \lambda^2}, & |\lambda| \le 2\sigma \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

implying eigenvalues are asymptotically confined to $[-2\sigma, 2\sigma]$, which sets a noise floor.

A related result applies to sample covariance matrices. Let $\mathbf{X} \in \mathbb{R}^{n\times p}$ have i.i.d. entries (zero mean, variance σ^2), and define $\mathbf{C} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$. As $n, p \to \infty$ with $p/n \to c > 0$, the ESD of $\mathbf{C}$ converges to the Marchenko-Pastur (MP) distribution (Yaskov 2016), supported on

$$[\sigma^2(1-\sqrt{c})^2,\ \sigma^2(1+\sqrt{c})^2]. \quad (2)$$

This bulk is the reference for detecting signal-bearing outlier eigenvalues.

Spiked Covariance Model and Signal Detection via Eigenvalue Outliers: The *spiked covariance model* (Paul 2007) describes data with a low-rank signal component $\mathbf{\Sigma}_{\text{signal}}$ of rank $k \ll p$ embedded in isotropic noise $\mathbf{\Sigma}_{\text{noise}} = \sigma^2\mathbf{I}_p$, so $\mathbf{\Sigma} = \mathbf{\Sigma}_{\text{signal}} + \mathbf{\Sigma}_{\text{noise}}$. If the signal is strong, top population eigenvalues λ_i' exceed the Baik-Ben Arous-Péché (BBP) threshold $\sigma^2(1+\sqrt{c})$, with $c = p/n$, and the corresponding sample eigenvalues of $\mathbf{C} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$ detach from the bulk (Yaskov 2016). While these RMT results are asymptotic, similar eigenvalue and eigenvector behavior is observed even for moderate p (e.g., $p = 768$–1024) (Benaych-Georges and Nadakuditi 2011). The BBP threshold is heuristically supported by our simulations (see our later deep-dive discussion in Sec. 4 showing robust subspace estimation even in finite samples). Such outlier eigenvalues/vectors enable projection onto signal-dominant directions, isolating semantics.

[1] This paper includes prompts and model outputs that may be offensive, used solely to illustrate adversarial vulnerabilities and guide the development of safer, more robust model defenses.

RMT-based Decomposition for Stable Subspace Extraction: To extract a subspace aligned with top spectral components, we approximate the empirical covariance matrix Σ_n of clean embeddings as a low-rank signal plus isotropic noise, fitting via

$$\mathcal{L}_{\mathrm{RMT}}(U, \Lambda) = \left\| \Sigma_n - (U \Lambda U^\top + \sigma^2 I_P) \right\|_F^2, \quad (3)$$

where $U \in \mathbb{R}^{P \times r}$ (orthonormal columns: top-r signal directions) and $\Lambda \in \mathbb{R}^{r \times r}$ (diagonal, corresponding eigenvalues). The subspace spanned by U is the *causal subspace*—a spectral subspace aligned with directions empirically robust to adversarial perturbations, as measured by the Robustness-based Nonconformity Score (RbNS). While this notion of causality is not interventional, it acts as a principled, architecture-agnostic heuristic for isolating directions that are empirically least susceptible to attack.

2.2 RMT-based Adversarial Noise Annihilation

Asymptotic Identifiability of the Causal Subspace

Theorem 1 *Let* $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_p + \sum_{k=1}^{K_C} \theta_{C,k} \mathbf{v}_{C,k} \mathbf{v}_{C,k}^\top + \sum_{j=1}^{K_A} \theta_{A,j} \mathbf{v}_{A,j} \mathbf{v}_{A,j}^\top$ *be the population covariance of embeddings, where* $\{\mathbf{v}_{C,k}\}$ *and* $\{\mathbf{v}_{A,j}\}$ *are orthonormal and correspond to* K_C *causal and* K_A *adversarial spikes.*

Let $\mathbf{S}_{adv}$ *be the sample covariance from* n *embeddings. Define the BBP threshold as* $\theta_{BBP} = \sigma^2(1 + \sqrt{y})^2$ *with* $p/n \to y > 0$. *Suppose all* $\theta_{C,k} > \theta_{BBP} + \eta_C$ *and all* $\theta_{A,j} < \theta_{BBP} - \eta_A$, *for some* $\eta_C, \eta_A > 0$. *Then the subspace* $\hat{\mathcal{S}}_C$, *spanned by the top* K_C *eigenvectors of* $\mathbf{S}_{adv}$, *satisfies*

$$\mathbf{P}_{\hat{\mathcal{S}}_C} \xrightarrow{a.s.} \mathbf{P}_{\mathcal{S}_C}, \quad \text{where } \mathcal{S}_C = \mathrm{span}(\{\mathbf{v}_{C,k}\}_{k=1}^{K_C}).$$

Claim 1 ***Adversarial Noise is Largely Orthogonal to the Causal Subspace:*** *Let* $\mathbf{e}$ *denote a clean embedding primarily supported in the causal subspace* $\mathcal{S}_C$. *A typical adversarial perturbation* $\boldsymbol{\delta}$ *(with* $\|\boldsymbol{\delta}\| \leq \epsilon$, *causing a model output change without altering semantic content) can be written as* $\boldsymbol{\delta} = \boldsymbol{\delta}_\perp + \boldsymbol{\delta}_\parallel$, *where* $\boldsymbol{\delta}_\perp \in \mathcal{S}_C^\perp$, $\boldsymbol{\delta}_\parallel \in \mathcal{S}_C$, *and typically* $\|\boldsymbol{\delta}_\parallel\| \ll \|\boldsymbol{\delta}_\perp\|$. *In favorable cases,* $\boldsymbol{\delta}_\parallel \approx \mathbf{0}$, *allowing near-complete removal via projection onto* $\mathcal{S}_C$.

Conditions for Asymptotic Orthogonality of Subspaces

Theorem 2 *Assume the conditions of Theorem 1 hold so that* $\hat{\mathcal{S}}_C \to \mathcal{S}_C$ *as* $p, n \to \infty$. *Suppose adversarial perturbations introduce distinct "adversarial spikes" spanning a subspace* $\mathcal{S}_A$, *with corresponding eigenvectors* $\{\mathbf{v}_{A,j}\}$, *and assume* $\mathcal{S}_C \perp \mathcal{S}_A$. *If (i) the causal and adversarial spikes exceed the BBP threshold and are well-separated in spectrum, and (ii) the subspace estimation mechanism correctly assigns sample eigenvectors to* $\hat{\mathcal{S}}_C$ *and* $\hat{\mathcal{S}}_A$, *then:*

$$\mathbf{P}_{\hat{\mathcal{S}}_C} \mathbf{P}_{\hat{\mathcal{S}}_A} \xrightarrow{a.s.} \mathbf{0} \quad \text{as } p, n \to \infty.$$

That is, the estimated causal and adversarial subspaces become asymptotically orthogonal.

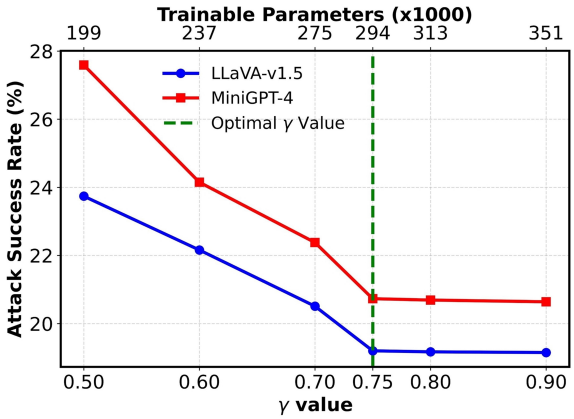

Figure 2: Attack Success Rate (%) vs. coverage parameter γ for LLaVA-v1.5 (Liu et al. 2024) (blue) and MiniGPT-4 (Zhu et al. 2023) (red). Higher γ retains more eigendirections, which corresponds to a lower attack success rate but increasing trainable parameter (top axis, in thousands).

Adversarial Noise Removal via Causal Projection

Theorem 3 *Let* $\mathbf{e}_{\mathrm{adv}} = \mathbf{e} + \boldsymbol{\delta}$ *be a perturbed embedding, with clean input* $\mathbf{e}$ *and adversarial perturbation* $\boldsymbol{\delta}$. *Assume: (i)* $\mathbf{P}_{\hat{\mathcal{S}}_C} \xrightarrow{a.s.} \mathbf{P}_{\mathcal{S}_C}$ *as* $p, n \to \infty$ *(Theorem 1), and (ii)* $\boldsymbol{\delta} \in \mathcal{S}_C^\perp$. *Then the projected embedding* $\mathbf{e}_{\mathrm{clean}} = \mathbf{P}_{\hat{\mathcal{S}}_C} \mathbf{e}_{\mathrm{adv}}$ *satisfies* $\|\mathbf{P}_{\hat{\mathcal{S}}_C} \boldsymbol{\delta}\| \xrightarrow{a.s.} 0$ *and* $\mathbf{e}_{\mathrm{clean}} \xrightarrow{a.s.} \mathbf{P}_{\mathcal{S}_C} \mathbf{e}$. *Thus, if* $\mathbf{e} \in \mathcal{S}_C$, *then* $\mathbf{e}_{\mathrm{clean}} \xrightarrow{a.s.} \mathbf{e}$.

Corollary 1 *For LLM embeddings that approximately conform to a spiked covariance model, i.e., semantic features align with dominant spectral directions, EigenShield can asymptotically suppress adversarial perturbations, provided these perturbations are largely orthogonal or spectrally separable from the signal subspace.*

Corollary 2 *For VLM embeddings (including unimodal components), if causal concepts yield identifiable spectral spikes and adversarial noise is not tightly coupled to these directions, projection-based filtering via EigenShield can achieve near-complete removal in the asymptotic limit.*

These guarantees are under idealized assumptions (e.g., spectral separability, low overlap between noise and signal). Practical effectiveness depends on how well real-world representations adhere to these conditions. Our later sections examine sensitivity to these assumptions.

3 Inference-time Defense with EigenShield

EigenShield operates in three steps (Fig. 1): *(i) Spectral Decomposition:* Compute the empirical covariance of input embeddings and extract principal directions via eigendecomposition. *(ii) RbNS-Based Thresholding:* Identify a causal subspace using a threshold set by the Robustness-based Nonconformity Score (RbNS); remaining components are treated as noise or adversarial. *(iii) Projection Filtering:* Project embeddings onto the causal subspace to

Defense	Attack Success	Identity Attack	Profanity	Sexually Explicit	Threat	Toxicity Success	Attack Attack	Identity	Profanity Explicit	Sexually	Threat	Toxicity
	GPT-2						Llama-3.1-8B-Instruct					
No Defense	82.34	32.53	61.22	45.81	38.15	70.53	55.62	14.23	32.81	20.14	12.48	38.92
LLaMA Guard	30.14	7.23	20.32	10.78	6.11	28.73	19.83	3.12	11.49	5.68	2.29	13.41
Erase-and-Check	25.53	5.82	16.51	8.49	4.17	22.48	16.52	2.47	9.13	4.48	1.69	10.19
EigenShield	**5.81**	**1.48**	**3.19**	**1.77**	**0.88**	**4.46**	**4.18**	**0.79**	**2.07**	**0.93**	**0.38**	**2.76**
	LLaVA-v1.5-7B						MiniGPT-4					
No Defense	81.61	25.41	67.22	39.38	40.64	77.93	37.20	2.94	26.53	12.76	2.10	31.57
Adv. Training	28.21	3.71	25.05	10.90	1.87	28.44	29.82	2.66	22.41	10.15	1.63	23.20
UNIGUARD	25.17	2.06	22.34	7.99	**0.86**	19.16	24.98	**1.37**	16.42	10.69	1.80	18.73
BLURKERNEL	39.03	3.92	30.61	14.10	3.17	32.28	38.92	2.28	28.34	13.79	2.12	33.08
COMP-DECOMP	37.70	2.67	29.02	13.26	3.59	31.94	35.21	2.31	25.56	11.97	1.54	29.06
DIFFPURE	40.42	3.01	30.89	14.48	3.35	34.06	41.32	2.12	29.89	15.24	2.12	35.65
CIDER	24.73	2.88	25.80	9.79	2.53	17.49	22.74	1.68	17.15	9.82	1.93	20.04
EigenShield	**19.20**	**1.76**	**15.31**	**5.16**	1.01	**13.28**	**20.37**	1.39	**12.80**	**8.53**	**0.96**	**15.77**

Table 1: Comparison of EigenShield and SOTA defenses on LLMs and VLMs. Metrics: attack success rate and content safety (lower is better). **Bold** = best, Underlined = second best.

suppress adversarial perturbations while preserving semantic structure. The approach is model-agnostic and does not modify the model. The methodology has two phases: *Phase 1:* Threshold selection using RbNS and RMT; *Phase 2:* Inference-time defense.

Phase 1: Threshold Selection via RbNS: Select the threshold τ^* to separate causal from correlational eigenvalues. RbNS quantifies robustness of spectral directions:

Outlier Extraction: Identify eigenvectors with eigenvalues above the MP bulk.

RbNS Computation: For each outlier, compute RbNS; lower scores indicate greater stability.

Quantile-Based Thresholding: Apply quantile thresholding on nonconformity scores to set τ^*. The coverage parameter γ determines the proportion of directions retained as causal. Fig. 2 shows performance trends for γ between 0.5 and 0.9; $\gamma = 0.75$ is used for an optimal trade-off. A direction is retained as causal if its RbNS score falls below the γ-th quantile; directions above are treated as noise or adversarial. Increasing γ improves coverage but increases complexity.

Phase 2: Inference-Time Jailbreak Defense: With τ^* from Phase 1, EigenShield defends LLMs and VLMs at inference. For each input, compute eigen-decomposition; directions with eigenvalues above τ^* are deemed causal. Project the embedding:

$$\mathbf{e}_{\text{filtered}} = \mathbf{P}_{\text{causal}}\mathbf{e}_{\text{input}}, \quad \mathbf{P}_{\text{causal}} = \mathbf{E}_{\text{causal}}\mathbf{E}_{\text{causal}}^{\top}. \quad (4)$$

The filtered embedding is used for downstream tasks, preserving semantics while suppressing adversarial effects. EigenShield requires no model modification.

Models. We evaluate EigenShield on a range of LLMs, GPT-2 (Radford et al. 2019), LLaMA-3.1-8B-Instruct (Grattafiori et al. 2024), Qwen-2.5-7B-Instruct (Yang et al. 2024), Gemma-7B (Team et al. 2024)—and VLMs, including LLaVA-v1.5-7B (Liu et al. 2024), MiniGPT-4 (Zhu et al. 2023), InstructBLIP (Dai et al. 2023), Qwen2-VL (Wang et al. 2024), and Florence-2-large (Xiao et al. 2023).

Datasets. LLMs are tested with custom jailbreak prompts and RealToxicityPrompts (Gehman et al. 2020). VLMs are evaluated on HarmBench (Mazeika et al. 2024), which includes adversarial PGD (Madry et al. 2017) examples, and on RealToxicityPrompts paired with adversarial images. For generalization, we construct a composite dataset spanning FGSM (Goodfellow, Shlens, and Szegedy 2014), PGD, MIM (Dong et al. 2018), CW (Carlini and Wagner 2017), and Square Attack (Andriushchenko et al. 2020).

Defense Baselines. LLM comparisons include: (i) No Defense, (ii) LLaMA-Guard (Inan et al. 2023), (iii) Erase-and-Check (Kumar et al. 2023). VLM baselines: (i) No Defense, (ii) image transformations (BlurKernel, Comp-Decomp), (iii) DIFFPURE (Nie et al. 2022), (iv) CIDER (Xu et al. 2024), (v) Adversarial Training (Shafahi et al. 2019), (vi) UNIGUARD (Oh et al. 2024).

Evaluation Metrics. We report (1) *Attack Success Rate (ASR)*, the fraction of adversarial inputs eliciting harmful or disallowed outputs. For VLMs:

$$\text{ASR} := \frac{1}{|\mathcal{D}|} \sum_{(\text{prompt}, x_{\text{adv}}) \in \mathcal{D}} \mathbb{I}_{\text{harm}}(\mathcal{G}(\mathcal{F}(\text{prompt}, x_{\text{adv}}))), \quad (5)$$

where $\mathcal{F}$ is the VLM, $\mathcal{G}$ is a Perspective API classifier, and $\mathbb{I}_{\text{harm}}$ indicates harmful output. (2) *Content Harmfulness Scores* from the Perspective API, toxicity, identity attack, profanity, sexually explicit, threat, each in $[0, 1]$.[2]

Robustness against Standard Jailbreaks. Table 1 shows EigenShield achieves the lowest ASR and content harm metrics across all models. On GPT-2, EigenShield reduces ASR from 82.34% (no defense) to 5.81%, outperforming LLaMA Guard and Erase-and-Check. Similar improvements hold for all safety metrics, including identity attack, profanity, sexual content, threat, and toxicity. On LLaVA-v1.5-7B, Eigen-

[2]**Limitations:** The Perspective API exhibits sociolinguistic and distributional biases, especially on marginalized dialects or OOD phrasing (Sap et al. 2022; Borkan et al. 2019). We use it for comparability with prior work and automated toxicity screening.

Metric	Llama-3.1		Qwen2.5		Gemma-7B		GPT-2		LLaVA-7B		MiniGPT-4		Qwen2-VL	
	Base	**ES**	Base	**ES**	Base	**ES**	Base	**ES**	Base	**ES**	Base	**ES**	Base	**ES**
Attack 1: GCG (Greedy Coordinate Gradient) (Zou et al. 2023)														
ASR	85.67	**16.92**	70.29	**12.08**	80.15	**15.33**	92.13	**20.54**	82.17	**21.09**	75.82	**18.56**	45.04	**10.81**
Identity	28.15	**3.11**	15.62	**1.93**	22.73	**2.76**	40.88	**5.87**	25.43	**4.17**	20.11	**3.04**	8.25	**1.28**
Profanity	60.43	**10.28**	45.88	**7.14**	55.09	**9.05**	75.19	**15.62**	58.02	**13.58**	50.69	**10.91**	25.93	**5.03**
Sexual	40.77	**5.16**	25.04	**3.28**	35.18	**4.59**	55.23	**8.11**	38.19	**7.12**	30.24	**5.37**	14.77	**2.16**
Threat	30.29	**3.54**	18.33	**1.99**	25.61	**2.81**	48.71	**6.03**	28.57	**4.03**	20.73	**2.98**	8.19	**1.04**
Toxicity	75.38	**18.05**	58.17	**13.24**	69.41	**16.57**	88.67	**22.19**	72.88	**22.74**	65.31	**19.62**	38.65	**11.49**
Attack 2: TAP (Tree of Attacks with Pruning) (Mehrotra et al. 2024)														
ASR	88.12	**19.65**	75.44	**15.21**	83.07	**18.04**	90.58	**23.18**	80.43	**23.71**	78.15	**20.98**	50.67	**13.26**
Identity	30.59	**4.28**	18.03	**2.87**	25.16	**3.91**	38.72	**6.94**	23.92	**5.08**	22.06	**4.13**	10.14	**2.07**
Profanity	63.78	**13.19**	50.19	**9.76**	58.61	**11.67**	72.33	**18.03**	55.78	**16.21**	53.11	**13.82**	30.27	**7.69**
Sexual	43.15	**7.02**	28.76	**5.03**	38.49	**6.14**	53.81	**9.86**	36.05	**8.74**	33.78	**7.06**	18.03	**4.02**
Threat	33.88	**4.93**	20.91	**3.08**	28.02	**4.11**	46.05	**7.15**	26.19	**5.22**	24.33	**4.17**	10.99	**2.21**
Toxicity	79.04	**21.73**	63.55	**16.81**	73.18	**19.36**	85.92	**25.49**	70.15	**25.03**	68.49	**22.19**	43.81	**15.04**
Attack 3: AutoBreach (Chen et al. 2024)														
ASR	89.53	**21.06**	78.19	**17.39**	86.74	**19.81**	93.22	**24.87**	85.37	**25.19**	80.91	**22.65**	55.83	**15.91**
Identity	32.18	**5.03**	20.47	**3.59**	28.33	**4.72**	42.05	**7.15**	28.04	**6.15**	24.77	**5.28**	13.06	**3.14**
Profanity	67.02	**15.21**	54.66	**11.94**	62.41	**13.85**	78.14	**19.88**	62.15	**18.77**	57.39	**16.04**	35.94	**9.88**
Sexual	46.93	**8.39**	31.85	**6.77**	42.06	**7.61**	58.39	**10.73**	42.33	**10.04**	37.16	**8.59**	22.11	**5.73**
Threat	36.75	**5.88**	24.38	**4.62**	31.92	**5.04**	50.17	**8.02**	32.08	**6.71**	28.41	**5.33**	14.82	**3.05**
Toxicity	82.19	**23.94**	68.70	**19.07**	77.58	**21.63**	90.03	**27.51**	76.92	**27.48**	72.03	**24.72**	49.07	**17.39**

Table 2: **EigenShield (ES) reduces attack success and content harm under adaptive jailbreaks.** Lower is better.

Metric	**GPT-2**		**Llama-3.1-8B**		**Qwen2.5**		**Gemma-7B**		**LLaVA-v1.5-7B**		**MiniGPT-4**		**Florence-2**		**InstructBLIP**	
	Base	**ES**	**Base**	**ES**	**Base**	**ES**	**Base**	**ES**	**Base**	**ES**	**Base**	**ES**	**Base**	**ES**	**Base**	**ES**
ASR	80.23	**10.51**	54.04	**8.13**	14.62	**4.91**	41.21	**7.34**	68.03	**14.29**	60.31	**11.88**	38.49	**8.63**	45.22	**10.04**
Identity	31.54	**2.92**	13.82	**1.84**	2.33	**0.67**	7.92	**1.41**	19.92	**3.86**	15.78	**2.74**	7.34	**1.42**	8.97	**1.66**
Profanity	60.17	**6.13**	31.93	**4.52**	7.91	**2.12**	19.73	**3.63**	50.27	**11.74**	44.16	**9.51**	25.06	**5.77**	30.11	**7.23**
Sexual	44.02	**3.84**	19.94	**2.51**	3.52	**0.83**	11.54	**1.82**	31.84	**6.91**	26.13	**5.03**	15.81	**3.59**	18.28	**4.09**
Threat	37.23	**1.91**	11.81	**0.96**	1.67	**0.37**	6.12	**0.73**	18.99	**2.83**	14.96	**2.27**	6.88	**1.24**	8.03	**1.51**
Toxicity	69.54	**9.22**	37.82	**6.34**	10.14	**3.02**	24.61	**5.23**	63.16	**15.82**	56.02	**13.06**	33.72	**8.38**	40.79	**11.17**

Table 3: **Results for OOD-style harmful inputs: Base (no defense) and EigenShield (ES, in bold)**. Lower is better.

Shield lowers ASR from 81.61% to 19.20%, outperforming both training-time (adversarial training, UNIGUARD) and inference-time (CIDER, DIFFPURE) defenses. Improvements extend to all harm dimensions, confirming the method's model-agnostic performance.

Defense against Adaptive Jailbreaks. Table 2 presents results for three state-of-the-art adaptive attacks: GCG, TAP, and AutoBreach. Across all LLMs and VLMs, EigenShield reduces ASR by 60–80 percentage points compared to the baseline. For example, under AutoBreach, Llama-3.1's ASR drops from 89.53% to 21.06%, and for LLaVA-7B from 85.37% to 25.19%. These reductions hold across all content harm metrics, including toxicity and sexual explicitness. This demonstrates that EigenShield remains effective even against strong, adaptive, white-box attacks.

Out-of-Distribution (OOD) Generalization. Table 3 evaluates EigenShield on OOD-style harmful inputs for four LLMs and four VLMs. Substantial ASR reductions are observed for all models, regardless of baseline robustness. For instance, in Qwen2.5, ASR drops from 14.62% to 4.91%; in MiniGPT-4, from 60.31% to 11.88%. Notably, EigenShield also reduces harm scores in challenging OOD regimes, indicating that it is not reliant on in-distribution artifacts.

Multimodal Robustness and Cross-Modal Generalization. EigenShield delivers substantial robustness gains against multimodal adversarial attacks, where text inputs are adversarial and images remain clean. Table 4 shows that EigenShield reduces attack success rates (ASR) by up to 63 points (LLaVA-v1.5-7B, RTP) and achieves large absolute reductions in all Perspective API harm metrics, such as up to 59.5 for toxicity and 51.7 for profanity. These improvements hold across a range of VLMs and both challenging text-based adversarial datasets (RTP and HarmBench). The defense generalizes to joint embedding spaces even when only one modality is attacked. EigenShield also provides meaningful gains for models with low baseline ASR or harm, such as Qwen2-VL and Florence-2-large. This confirms its effectiveness for both high-risk and low-error scenarios.

Efficacy on Clean Data. Table 5 reports absolute reductions in toxicity metrics on clean text-image pairs. While gains

Model	ASR	Identity	Profanity	Sex.	Threat	Toxicity
RTP (Gehman et al. 2020)						
LLaVA-v1.5-7B	63.1	20.3	51.7	34.4	43.1	59.5
MiniGPT-4	36.8	3.1	24.0	11.8	3.6	31.1
InstructBLIP	25.0	4.4	20.7	13.5	3.9	24.1
Qwen2-VL	15.2	1.9	10.1	4.8	2.4	12.3
Florence-2-large	24.8	3.4	17.6	7.6	3.7	20.5
HarmBench (Mazeika et al. 2024)						
LLaVA-v1.5-7B	47.3	13.7	35.2	23.1	26.0	45.2
MiniGPT-4	32.3	4.2	20.1	11.0	4.1	26.3
InstructBLIP	10.6	2.1	6.8	3.7	1.5	9.0
Qwen2-VL	11.1	1.4	7.5	3.3	1.4	8.7
Florence-2-large	13.6	2.6	8.5	4.8	2.5	11.8

Table 4: EigenShield improvement (Base – ES) under multimodal attacks: adversarial text (RTP, HB) with clean images. Metrics are absolute reduction in ASR and Perspective API scores; higher is better.

LLMs			VLMs		
Model	RTP	HB	**Model**	RTP	HB
GPT-2	2.0	1.5	LLaVA	0.8	0.5
Gemma	0.6	0.4	MiniGPT	0.5	0.3
Llama-3	0.4	0.4	InstructBLIP	0.3	0.4
Qwen2.5	0.3	0.1	Qwen2-VL	0.0	0.3

Table 5: **EigenShield improvement (Base – ES) on toxicity scores for LLMs and VLMs on clean (RTP, HB) data.** Values are absolute reductions; higher is better.

	LLMs		VLMs	
Defense	**GPT-2**	**Llama-3**	**LLaVA**	**MiniGPT-4**
Baseline	181.4	403.8	452.1	473.5
EigenShield	**190.6**	**480.1**	**466.2**	**487.9**
LLaMA Guard	532.6	755.0	–	–
Erase-and-Check	454.9	1008.9	–	–
Adversarial Training*	–	–	452.1	473.5
UNIGUARD*	–	–	452.1	473.5
BLURKERNEL	–	–	458.4	479.8
COMP-DECOMP	–	–	458.4	479.8
CIDER	–	–	513.9	535.3
DIFFPURE	–	–	1667.8	1689.2

Table 6: Latency Comparisons (ms)

*Training-time method (no inference overhead). '–' indicates N/A.

are naturally smaller in this regime, EigenShield still delivers measurable improvements, especially for models with higher initial risk like GPT-2 and LLaVA. At the same time, it avoids over-filtering and does not degrade benign semantic content, confirming that the method selectively targets adversarial spectral anomalies.

Runtime Analysis. Table 6 shows that EigenShield is suitable for real-time use. It introduces minimal latency (9.2–14.4 ms) for LLMs and VLMs, significantly outperforming methods like LLaMA Guard and DIFFPURE, which multiply inference times. While basic pre-processing is faster, EigenShield offers a superior, model-agnostic defense without the high computational cost of other methods.

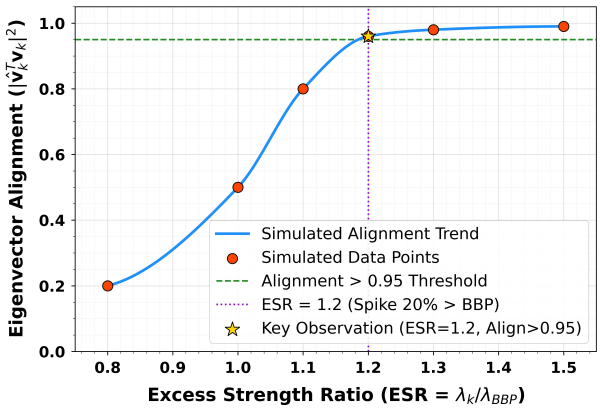

Figure 3: Eigenvector alignment ($|\hat{\mathbf{v}}_k^T \mathbf{v}_k|^2$) versus Excess Strength Ratio (ESR). Phase transition near ESR$= 1$, with alignment > 0.95 for ESR$\geq$1.2 ($\lambda_k > 20\%$ above bulk).

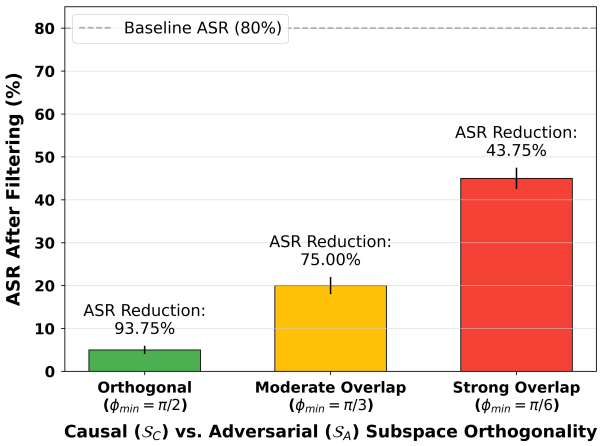

Figure 4: ASR after EigenShield filtering as a function of the principal angle ϕ_{min} between causal ($\mathcal{S}_C$) and adversarial ($\mathcal{S}_A$) subspaces. Baseline ASR is 80% (dashed line). As ϕ_{min} decreases from $\pi/2$ (orthogonal) to $\pi/6$ (strong overlap), post-filtering ASR increases, reflecting diminished defense. Error bars: ± 1 standard deviation.

4 Analysis and Discussion

Although EigenShield shows strong empirical robustness, its theoretical guarantees rely on the asymptotic regime of RMT, where embedding dimension p and sample size n are large ($p/n \to y > 0$). In this regime, projection onto the estimated causal subspace provably removes adversarial noise orthogonal to the signal while preserving semantics. These guarantees depend on two conditions: (i) the estimated subspace converges to the true signal subspace as $p, n \to \infty$, and (ii) adversarial perturbations are largely orthogonal to this subspace. In practice, these ideal conditions are rarely met exactly. However, as shown previously, EigenShield remains highly effective even in finite-sample and non-ideal regimes. We further analyze why its empirical robustness persists and clarify the limits of its practical effectiveness.

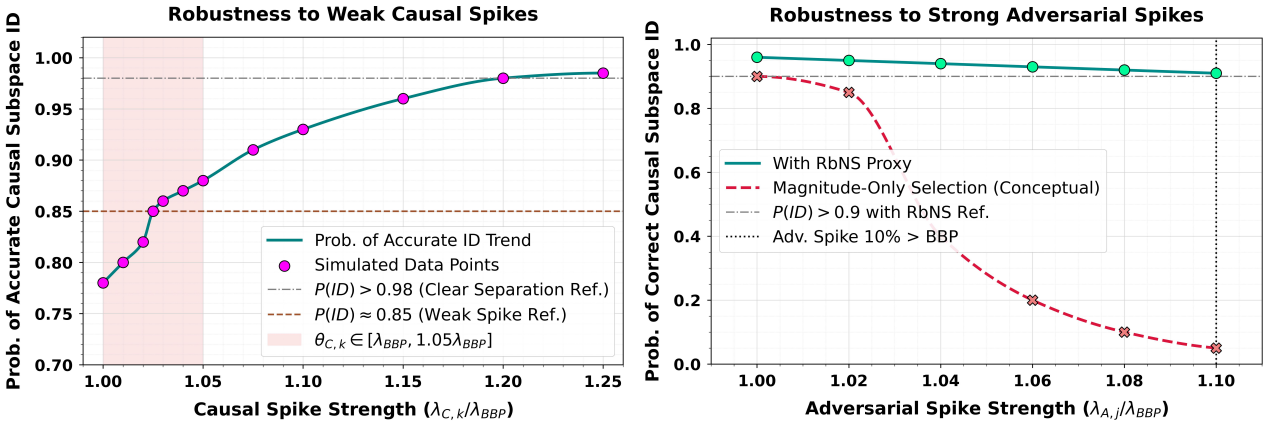

Figure 5: Simulated robustness of causal subspace identification. **(a) Weak Causal Spikes:** Identification accuracy for $\hat{\mathcal{S}}_C$ remains high ($\sim$0.85) even when causal spike strength $\lambda_{C,k}$ is only 5% above the BBP threshold. **(b) Strong Adversarial Spikes:** The RbNS proxy (dark cyan) maintains > 0.9 accuracy against rising adversarial spike strength $\lambda_{A,j}$, significantly outperforming magnitude-only selection (dashed crimson) for spikes up to 10% above λ_{BBP}.

Finite-Sample Behavior In real applications, both p and n are finite, p typically ranges from 768 to 4096 for modern LLMs and VLMs, while n is limited by batch or dataset size. To test the applicability of RMT-based guarantees, we simulated spiked covariance models across a range of spike strengths (θ_k) and noise variances (σ^2).

Fig. 3 depicts how the recovery of population eigenvectors depends on the Excess Strength Ratio (ESR), defined as λ_k/λ_+, where $\lambda_k = \sigma^2 + \theta_k$ is the population signal eigenvalue, and $\lambda_+ = \sigma^2(1+\sqrt{c})^2$ is the theoretical upper edge of the noise bulk. When ESR ≤ 1, eigenvector alignment ($|\hat{\mathbf{v}}_k^T \mathbf{v}_k|^2$) is poor, indicating that signal cannot be reliably distinguished from noise. As ESR increases above 1, alignment improves rapidly: for ESR $= 1.2$, recovery is already above 0.95; for ESR ≥ 1.3, it approaches unity. This transition explains why EigenShield works robustly in practice: real-world embeddings often contain dominant spectral directions strong enough to be reliably separated from noise, even with moderate sample sizes.

Dependence on Subspace Overlap The effectiveness of subspace-projection defenses like EigenShield is fundamentally limited by the overlap between the causal subspace ($\mathcal{S}_C$) and the subspace spanned by adversarial perturbations ($\mathcal{S}_A$). We systematically varied the smallest principal angle ϕ_{min} between $\mathcal{S}_C$ and $\mathcal{S}_A$ in synthetic experiments, reporting the post-filtering Attack Success Rate (ASR):

- **Orthogonal subspaces** ($\phi_{min} = \pi/2$): EigenShield reduces baseline ASR from 80% to 5% ($\sim$ 94% reduction).
- **Moderate overlap** ($\phi_{min} = \pi/3$): ASR after filtering rises to 20%.
- **Strong overlap** ($\phi_{min} = \pi/6$): ASR rises to 45%.

As illustrated in Fig. 4, the defense is highly effective when adversarial directions are largely orthogonal to the signal manifold, but its efficacy drops rapidly as adversarial perturbations become more aligned with the causal subspace. In the extreme, no spectral projection method can eliminate adversarial components that are indistinguishable from the true semantic directions the model must preserve.

Robustness to Weak Causal and Strong Adversarial Spikes Two common non-idealities are (a) causal spikes that are only marginally stronger than noise, and (b) adversarial spikes that are both strong and variable. Fig. 5(a) shows that even when the causal spike strength is within 5% of the noise threshold, the probability of correct subspace identification remains high ($\sim$0.85), and rises sharply for stronger signals (> 0.98 when the spike is 20% above noise). This demonstrates a smooth, not catastrophic, degradation in performance. Fig. 5(b) analyzes robustness when adversarial spikes approach or exceed the bulk edge. If these directions are unstable (i.e., high RbNS), EigenShield's reliance on both spectral magnitude and stability allows it to maintain > 0.9 probability of correct subspace selection. In contrast, selection based on eigenvalue magnitude is easily defeated by spectrally strong but unstable components.

5 Conclusions

We presented EigenShield, a principled, inference-time, model-agnostic defense for LLMs and VLMs, requiring no model access or retraining. By leveraging RMT and RbNS, it reliably identifies and preserves causal subspaces while suppressing adversarial components. Across diverse models, datasets, and attack types, including adaptive, OOD, and multimodal, EigenShield consistently reduces attack success and content harm. Its effectiveness persists even as RMT assumptions are relaxed, with performance degrading gracefully in finite-sample or imperfect spectral regimes. The method outperforms state-of-the-art adaptive and multimodal defenses, reducing attack success by up to 93% for LLMs and 80% for VLMs. EigenShield is computationally efficient, adding only minimal latency. However, like all projection-based methods, it cannot defend against attacks aligned with the preserved causal manifold.

Broader Impact

EigenShield is an inference-time, model-agnostic defense that suppresses adversarial perturbations in LLMs and VLMs by projecting embeddings onto a spectrally inferred causal subspace. Its practical relevance lies in reducing jailbreaking-induced content generation, without requiring access to model weights or retraining.

Positive Impact. By reducing Attack Success Rates by up to 92.9% for LLMs and 76.5% for VLMs, EigenShield directly limits harmful content emission. Its plug-in nature makes it deployable across a wide range of black-box APIs and production models. The causal filtering approach leverages signal-level structure–without modifying model internals–offering a new class of defenses rooted in spectral statistics. This architecture-agnostic robustness is essential for securing frontier models in real-world deployment, especially in domains where retraining is infeasible.

Risks. EigenShield's robustness depends on spectral separability of causal and adversarial directions. Under critical alignment (Appendix A.8), this assumption may fail, leading to partial or no filtering. Moreover, hyperparameter tuning (e.g., γ) must trade off ASR reduction and model utility; over-filtering can degrade performance on benign OOD inputs.

Additionally, the causal subspace is learned from data. If this data contains spurious correlations or biases, the filtered subspace may encode and preserve them. This could disproportionately affect inputs from underrepresented groups, especially in multilingual or non-Western contexts, where spectral profiles diverge from dominant training corpora. Finally, the defense may incentivize stronger adaptive attacks specifically designed to evade RMT-based projections. Since EigenShield's RbNS filtering is non-transparent and sensitive to spectral mass distribution, diagnosing failure cases and ensuring auditability remain open problems.

Mitigation. Future work should explore adaptive causal subspace tracking under distribution shift, improve interpretability of RbNS decisions, and audit for demographic bias in causal eigenstructure. Transparent reporting of EigenShield's limitations is critical for its safe use in high-stakes settings.

Acknowledgements

This work was supported in part by COGNISENSE, one of seven centers in JUMP 2.0, a Semiconductor Research Corporation (SRC) program sponsored by DARPA and by NSF Award #2046435.

References

Andriushchenko, M.; Croce, F.; Flammarion, N.; and Hein, M. 2020. Square attack: a query-efficient black-box adversarial attack via random search. In *European conference on computer vision*, 484–501. Springer.

Bao, Z.; Wang, Y.; Wang, Y.; and Xu, Z. 2022. Eigenvalue spectrum of deep neural network covariance matrices: Spiked random matrix theory and beyond. *Journal of Machine Learning Research*, 23(1): 1–52.

Benaych-Georges, F.; and Nadakuditi, R. R. 2011. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1): 494–521.

Borkan, D.; Dixon, L.; Sorensen, J.; Thain, N.; and Vasserman, L. 2019. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 330–336.

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, 39–57. Ieee.

Chan, T. 1992. The Wigner semi-circle law and eigenvalues of matrix-valued diffusions. *Probability theory and related fields*, 93(2): 249–272.

Chen, J.; Yang, X.; Fang, Z.; Tian, Y.; Dong, Y.; Yin, Z.; and Su, H. 2024. Autobreach: Universal and adaptive jailbreaking with efficient wordplay-guided optimization. *arXiv preprint arXiv:2405.19668*.

Dai, W.; Li, J.; LI, D.; Tiong, A.; Zhao, J.; Wang, W.; Li, B.; Fung, P. N.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 49250–49267. Curran Associates, Inc.

Das, N.; Shanbhogue, M.; Chen, S.-T.; Hohman, F.; Chen, L.; Kounavis, M. E.; and Chau, D. H. 2017. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression. *arXiv preprint arXiv:1705.02900*.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9185–9193.

Fischer, T.; Biemann, C.; et al. 2024. Large language models are overparameterized text encoders. *arXiv preprint arXiv:2410.14578*.

Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; and Smith, N. A. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Inan, H.; Upasani, K.; Chi, J.; Rungta, R.; Iyer, K.; Mao, Y.; Tontchev, M.; Hu, Q.; Fuller, B.; Testuggine, D.; et al. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*.

Jain, A.; Zhang, Y.; Bamboo, K.; Lee, B.; Goldblum, M.; Geiping, J.; and Goldstein, T. 2023. Baselines for Quantifying and Improving the Robustness of Generative Language Models. *arXiv preprint arXiv:2302.03293*.

Jin, H.; Hu, L.; Li, X.; Zhang, P.; Chen, C.; Zhuang, J.; and Wang, H. 2024. Jailbreakzoo: Survey, landscapes, and horizons in jailbreaking large language and vision-language models. *arXiv preprint arXiv:2407.01599*.

Kaneko, T.; and Harada, T. 2021. Blur, noise, and compression robust generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13579–13589.

Kumar, A.; Agarwal, C.; Srinivas, S.; Li, A. J.; Feizi, S.; and Lakkaraju, H. 2023. Certifying llm safety against adversarial prompting. *arXiv preprint arXiv:2309.02705*.

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

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Mazeika, M.; Phan, L.; Yin, X.; Zou, A.; Wang, Z.; Mu, N.; Sakhaee, E.; Li, N.; Basart, S.; Li, B.; et al. 2024. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*.

Mehrotra, A.; Zampetakis, M.; Kassianik, P.; Nelson, B.; Anderson, H.; Singer, Y.; and Karbasi, A. 2024. Tree of attacks: Jailbreaking black-box llms automatically. *Advances in Neural Information Processing Systems*, 37: 61065–61105.

Nie, W.; Guo, B.; Huang, Y.; Xiao, C.; Vahdat, A.; and Anandkumar, A. 2022. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*.

Oh, S.; Jin, Y.; Sharma, M.; Kim, D.; Ma, E.; Verma, G.; and Kumar, S. 2024. Uniguard: Towards universal safety guardrails for jailbreak attacks on multimodal large language models. *arXiv preprint arXiv:2411.01703*.

Paul, D. 2007. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 1617–1642.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Sap, M.; Gabriel, S.; Qin, L.; Smith, N. A.; and Choi, Y. 2022. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5879–5895. Association for Computational Linguistics.

Shafahi, A.; Najibi, M.; Ghiasi, M. A.; Xu, Z.; Dickerson, J.; Studer, C.; Davis, L. S.; Taylor, G.; and Goldstein, T. 2019. Adversarial training for free! *Advances in neural information processing systems*, 32.

Shayegani, E.; Dong, Y.; and Abu-Ghazaleh, N. 2023. Jailbreak in pieces: Compositional adversarial attacks on multimodal language models. In *The Twelfth International Conference on Learning Representations*.

Team, G.; Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M. S.; Love, J.; et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Vrbanec, T.; and Meštrović, A. 2020. Corpus-based paraphrase detection experiments and review. *Information*, 11(5): 241.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Fan, Y.; Dang, K.; Du, M.; Ren, X.; Men, R.; Liu, D.; Zhou, C.; Zhou, J.; and Lin, J. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*.

Weidinger, L.; Mellor, J.; Rauh, M.; Griffin, C.; Uesato, J.; Huang, P.-S.; Cheng, M.; Glaese, M.; Balle, B.; Kasirzadeh, A.; et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Xiao, B.; Wu, H.; Xu, W.; Dai, X.; Hu, H.; Lu, Y.; Zeng, M.; Liu, C.; and Yuan, L. 2023. Florence-2: Advancing a unified representation for a variety of vision tasks. *arXiv preprint arXiv:2311.06242*.

Xu, Y.; Qi, X.; Qin, Z.; and Wang, W. 2024. Defending jailbreak attack in vlms via cross-modality information detector. *arXiv e-prints*, arXiv–2407.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Yaskov, P. 2016. A short proof of the Marchenko–Pastur theorem. *Comptes Rendus Mathematique*, 354(3): 319–322.

Zhang, X.; Zhang, C.; Li, T.; Huang, Y.; Jia, X.; Xie, X.; Liu, Y.; and Shen, C. 2023. A mutation-based method for multi-modal jailbreaking attack detection. *arXiv preprint arXiv:2312.10766*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Zou, A.; Wang, Z.; Carlini, N.; Nasr, M.; Kolter, J. Z.; and Fredrikson, M. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

Zumbach, G. 2011. Empirical properties of large covariance matrices. *Quantitative Finance*, 11(7): 1091–1102.