Rethinking Jailbreak Detection of Large Vision Language Models with Representational Contrastive Scoring

Peichun Hua[1] **Hao Li**[1] **Shanghao Shi**[1] **Zhiyuan Yu**[2] **Ning Zhang**[1]
peichunhua04@gmail.com {li.hao, shanghao, zhang.ning}@wustl.edu
zhiyuanyu@tamu.edu
[1]Washington University in St. Louis [2]Texas A&M University

Abstract

Large Vision-Language Models (LVLMs) are vulnerable to a growing array of multimodal jailbreak attacks, necessitating defenses that are both generalizable to novel threats and efficient for practical deployment. Many current strategies fall short, either targeting specific attack patterns, which limits generalization, or imposing high computational overhead. While lightweight anomaly-detection methods offer a promising direction, we find that their common one-class design tends to confuse unseen benign inputs with malicious ones, leading to unreliable over-rejection. To address this, we propose Representational Contrastive Scoring (RCS), a framework built on a key insight: the most potent safety signals reside within the LVLM's own internal representations. Our approach inspects the internal geometry of these representations, learning a lightweight projection to maximally separate benign and malicious inputs in safety-critical layers. This enables a simple yet powerful contrastive score that differentiates true malicious intent from mere distribution shift. Our instantiations, MCD (Mahalanobis Contrastive Detection) and KCD (K-nearest Contrastive Detection), achieve state-of-the-art performance on a challenging evaluation protocol designed to test generalization to unseen attack types. This work demonstrates that effective jailbreak detection can be achieved by applying simple, interpretable statistical methods to the internal representations, offering a practical path towards safer LVLM deployment. Our code is available on Github[1].

1 Introduction

Large Vision-Language Models (LVLMs) — ranging from leading proprietary systems like GPT-4o and Gemini 2.5 Pro to powerful open-weight architectures such as LLaVA (Liu et al., 2023b, 2024a), Qwen3-VL (Bai et al., 2025a), Gemma 3 (Kamath et al., 2025), and InternVL3 (Zhu et al., 2025b) — have revolutionized multimodal AI capabilities, enabling sophisticated reasoning across text and visual inputs. However, this expanded capability introduces new vulnerabilities. Attackers can now exploit multiple modalities to bypass safety mechanisms through adversarial images (Qi et al., 2024a; Jeong et al., 2025), cross-modal prompt injection (Gong et al., 2025b), and traditional text-based jailbreaks transferred to LVLMs (Luo et al., 2024; Liu et al., 2024b). These diverse attack vectors pose significant challenges for deploying LVLMs safely in real-world applications.

A primary challenge in the safe deployment of LVLMs is developing defenses that are both **generalizable** to unseen attacks and **efficient** for real-time use. Existing strategies often compromise on these fronts: alignment-based methods and input filters (Liu et al., 2024g; Zong et al., 2024) tend to overfit to known attack patterns, leaving models vulnerable to emerging threats. Conversely, detection frameworks relying on consistency checks, gradients, or multiple inferences (Zhang et al., 2023; Xie et al., 2024; Wang et al., 2024a) often impose prohibitive computational overheads. This dichotomy between brittle specificity and high latency necessitates a shift toward more universal, lightweight frameworks.

Recently, JailDAM (Nian et al., 2025) was proposed as a more promising and efficient direction, framing jailbreak detection as an anomaly or Out-of-Distribution (OOD) problem. By learning to model the distribution of normal, benign inputs, these methods can detect deviations without needing to be trained on specific attack samples. However, our investigation reveals that when trained exclusively on benign data, these models tend to confuse *malicious intent* with *mere distribution shift*. Consequently, they suffer from a high rate of over-refusal, incorrectly flagging legitimate but

[1]https://github.com/sarendis56/Jailbreak_Detection_RCS

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 21748–21785

unseen benign prompts as harmful, which limits their reliability in diverse open-world settings. This highlights the need for a method that retains efficiency and generality while explicitly differentiating distribution shift from true maliciousness.

To address these challenges, we propose **Representational Contrastive Scoring (RCS)**. Our core intuition is that the most potent safety signals are not found in general-purpose embeddings, such as CLIP (Radford et al., 2021) used by JailDAM (Nian et al., 2025), but are encoded within the target model's own intermediate representations as it processes a prompt. Recent work in representation engineering (Zhou et al., 2024b; He et al., 2025) supports this, demonstrating that specific layers within LLMs reveal distinct geometric signatures for malicious versus benign inputs. Motivated by this and the concept of outlier exposure (Du et al., 2022; Hendrycks et al., 2019) from the OOD literature, RCS is designed to efficiently find and leverage these internal geometric signatures. Our lightweight framework operates with three key phases: (1) pinpointing the most discriminative hidden layers through a principled geometric analysis, (2) learning a lightweight projection that amplifies safety-relevant signals, and (3) scoring inputs based on their relative distance to benign vs. malicious samples in this projected space.

We instantiate this framework with two methods: **Mahalanobis Contrastive Detection (MCD)** and **K-nearest Contrastive Detection (KCD)**. Our comprehensive experiments show that they consistently outperform strong baselines on a challenging evaluation protocol that mixes data sources and modalities, while remaining both lightweight and flexible for real-world applicability.

2 Preliminaries

2.1 Jailbreak Attacks and Defense

Our study is closely related to LVLM safety. Below, we present a review of recent attacks and defenses in this domain, while deferring a more comprehensive review to Section A.

Jailbreak Attacks Against LVLMs. Jailbreak attacks against LVLMs have evolved into sophisticated multimodal strategies. **Text-based attacks** include gradient-based optimization methods (Zou et al., 2023b), role-playing (Li et al., 2023b), and multi-turn conversational exploits that gradually escalate malicious intent (Russinovich et al., 2025; Ren et al., 2025; Xiong et al., 2026). **Visual attacks** exploit the vision component through adversarial images (Qi et al., 2024a; Liu et al., 2023a), out-of-distribution visual inputs that fool safety guardrails (Jeong et al., 2025), typographical attacks (Gong et al., 2025b; Liu et al., 2024d), or by embedding hidden instructions in images (Schlarmann and Hein, 2023).

Limitations of Other Defense Paradigms. Despite extensive research, existing defenses suffer from fundamental limitations that hinder real-world deployment. **Safety Alignment** (Liu et al., 2024g; Zhang et al., 2025b) requires extensive retraining with curated multimodal datasets and substantial computational resources; yet, it remains fragile to unseen attack strategies (Yi et al., 2024; Qi et al., 2025). **Input filters** and **output classifiers** (Han et al., 2024; Chi et al., 2024) typically target specific attacks, failing to generalize to emerging threats. Meanwhile, some of them employ external large language models (guard models) to judge the safety of the conversation (Zhu et al., 2025a; Liu et al., 2025c), bringing significant memory and latency overhead, especially when reasoning is enabled on the guard models. Methods like MMCert (Wang et al., 2024a), GradSafe (Xie et al., 2024), and JailGuard (Zhang et al., 2023) require multiple model inferences or gradient computations, making them impractical for high-throughput applications.

Representation Engineering for Safety. Recent work demonstrates that LLM intermediate representations encode rich semantic information about input intent and safety (Arditi et al., 2024; Zhou et al., 2024b). Studies show that specific layers correlate with harmful content generation (Wu et al., 2024; He et al., 2025) and demonstrate empirical separability between benign and malicious prompts in representation space (Zhou et al., 2024b; He et al., 2024; Zhao et al., 2025a). However, these approaches rely on simple prototype classifiers on raw embeddings and limit their scope to text-only models. We systematically extend this paradigm to the multimodal domain, introducing a principled method to identify safety-critical layers and model their distributional geometry for robust detection across diverse modalities.

2.2 Problem Formulation

We formalize jailbreak detection as an out-of-distribution recognition problem that can utilize both benign and malicious training samples. Let $\mathcal{X}$ denote the space of all possible prompts (text, images, or multimodal combinations), and let f :

$\mathcal{X} \to \mathbb{R}^d$ represent a feature extractor that maps prompts to d-dimensional representations derived from the intermediate layers of the target LVLM.

We assume access to **benign training data**: $\mathcal{D}_{\text{benign}} = \{x_i\}_{i=1}^{N_b}$, where $x_i \sim P_{\text{benign}}(\mathcal{X})$, and **malicious training data**: $\mathcal{D}_{\text{malicious}} = \{x_j\}_{j=1}^{N_m}$, where $x_j \sim P_{\text{malicious}}(\mathcal{X})$. Both datasets are drawn from diverse sources to capture the heterogeneity of real-world usage.

Our objective is to design a detector $\delta : \mathcal{X} \to \{0, 1\}$ that: 1) Achieves **high detection performance** and **low false positives** across diverse benign prompts and jailbreak attempts (both text-only and multimodal attacks); 2) **Generalizes** to novel attack strategies not seen during training; 3) Operates **efficiently** at inference time without requiring post-training tuning, gradient computation, or multiple inferences. In particular, our detection methodology is capable of *making reliable decisions before decoding*, thereby reducing the inference expenses associated with malicious prompts.

3 Proposed Approach

3.1 Overview

Our **Representational Contrastive Scoring (RCS)** adapts classical OOD detection principles to utilize both benign and malicious data for robust jailbreak detection. We propose two distinct instantiations: MCD (Mahalanobis Contrastive Detection) and KCD (K-nearest Contrastive Detection). Both share a foundational process: (1) Principled Layer Selection via geometric analysis, and (2) Feature Extraction through a safety-aware learned projection. From this safety-aware representation space, MCD parametrically models the benign and malicious classes as sets of Gaussian distributions to perform scoring, whereas KCD non-parametrically scores inputs based on their relative distance to the k-nearest benign and malicious neighbors. By succeeding with both, we demonstrate that the effectiveness of our framework is not tied to specific distributional assumptions.

3.2 Identifying Safety-Critical Layers via Geometric Analysis

The effectiveness of representation-based detection hinges on identifying which layers encode the most discriminative safety signals. While prior work has empirically noted the utility of shallow or middle layers for text-based jailbreak detection (Zhao et al., 2025a), such ad-hoc selections do not easily scale to the complex, multimodal representations of LVLMs. To quantify this separability and pinpoint the optimal layers, we propose a principled, data-driven methodology. Our approach is founded on a central hypothesis: the layers with the highest downstream detection performance are those where *benign and malicious prompt representations are most geometrically separable*. An overview is depicted in Figure 1.

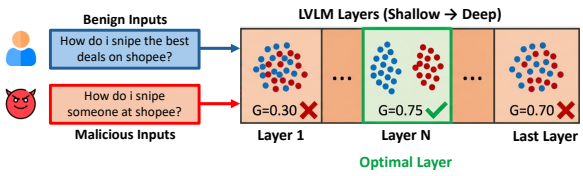


Figure 1: Layer selection by identifying safety-critical layers (Section 3.2)

To quantify this separability, we use the SGXSTest dataset (Gupta et al., 2024), which consists of carefully constructed pairs of benign and malicious prompts that are semantically almost identical. This paired design provides a controlled comparison, ensuring that any measured geometric separation is attributable to safety-relevant distinctions rather than spurious topic or style variations. This provides a clean signal for identifying truly discriminative layers. We compute a composite score for each layer based on three complementary metrics using the last-token representations: 1) **Maximum Margin Separation** ($\gamma^{(l)}$): We use a linear Support Vector Machine (SVM) to measure the width of the decision boundary. A wider margin indicates better linear separability and a stronger generalization potential. 2) **Cluster Cohesion** ($\mathcal{S}^{(l)}$): We employ the Silhouette Score to quantify how dense and well-separated the clusters are. For a sample i, $s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))}$, where $a(i)$ is the mean intra-cluster distance and $b(i)$ is the mean distance to the nearest foreign cluster. 3) **Discriminative Ratio** ($\mathcal{R}^{(l)}$): We compute the ratio of inter-class distance to pooled intra-class variance: $\mathcal{R}^{(l)} = \frac{\|\mu_{\text{benign}} - \mu_{\text{malicious}}\|_2}{\frac{1}{2}(\sigma_{\text{benign}} + \sigma_{\text{malicious}})}$, where σ denotes the average distance of samples to their respective centroids. A higher ratio signifies that the two distributions are far apart relative to their internal variance.

Our analysis, detailed in Section D, consistently identifies a representational "sweet spot" in the middle layers, which achieves higher performance in jailbreak detection (Section D.5). This finding

aligns with established representation learning theory (Zou et al., 2023a) where early layers capture low-level features, while final layers are often too specialized for the pretraining objective; the middle layers, in contrast, encode the rich, high-level semantic abstractions necessary to distinguish subtle malicious intent from benign queries. We further verify the robustness of this methodology in Section D.6, demonstrating that the optimal layer "sweet spot" can be reliably identified even with noisier, unpaired datasets, ensuring applicability in real-world scenarios where high-quality paired data may be scarce.

3.3 Feature Extraction and Safety-Aware Projection

Following layer selection, we extract the hidden state of the last token at the optimal layer:

$$f(x) = H^{(l^*)}(x)_{\text{last}} \in \mathbb{R}^{d_{\text{model}}} \tag{1}$$

where l^* denotes the selected layer. This position is critical as it represents the model's state immediately before generation, capturing the aggregated context of system prompts and user queries, making it an effective point for identifying malicious intent. At the same time, extracting features *before* response generation allows us to detect jailbreak attempts proactively without requiring the model to generate potentially harmful content. We confirm in Section C.4 that the last-token extraction strategy achieves superior performance compared to both mean pooling and extracting the last 5 tokens.

However, raw LVLM features pose practical challenges to training the detector: (i) high dimensionality (e.g., 4096) leads to the curse of dimensionality, where both kNN search and covariance estimation, essential to our methods discussed later in Section 3.5 and Section 3.4, become unstable given a limited number of training samples (Pestov, 2013; Ledoit and Wolf, 2004; Chen et al., 2011); and (ii) many dimensions encode task-irrelevant information, leading to suboptimal performance of our detectors without proper feature engineering. We address these through a learned neural projection $g_\theta : \mathbb{R}^{d_{\text{model}}} \to \mathbb{R}^{d_{\text{proj}}}$ (where $d_{\text{proj}} = 256$) shown in Figure 2, optimized for two objectives:

Dataset Clustering: Samples from the same dataset should cluster together while different datasets remain separated, preserving the natural structure of diverse benign sources. Here, m_d represents a margin hyperparameter, and only distances above it between two samples from different datasets will be penalized.

$$\begin{aligned}\mathcal{L}_{\text{dataset}} = &\sum_{d_i=d_j} \|g_\theta(x_i) - g_\theta(x_j)\|_2 \\ &+ \sum_{d_i\neq d_j} \max(0, m_d - \|g_\theta(x_i) - g_\theta(x_j)\|_2)\end{aligned} \tag{2}$$

Safety Separation: This term maximally separates the benign and malicious distributions. We take μ as the centroid of the corresponding cluster:

$$\mathcal{L}_{\text{sep}} = \max(0, m_s - \|\mu_{\text{benign}} - \mu_{\text{malicious}}\|_2) \tag{3}$$

The combined objective $\mathcal{L} = \alpha\mathcal{L}_{\text{dataset}} + \beta\mathcal{L}_{\text{sep}}$ is optimized using a three-layer feedforward network with batch normalization and dropout. This projection amplifies safety-relevant signals while suppressing irrelevant variations, ensuring that unseen benign inputs remain geometrically distinct from malicious clusters. We show in Section C.4 that this projection strategy is essential to performance, surpassing the baselines where the high-dimensional feature is not projected or reduced to low dimensions with PCA (Abdi and Williams, 2010).

3.4 Mahalanobis Contrastive Detection (MCD)

MCD models both benign and malicious distributions parametrically. Given the heterogeneity of real-world data, we model each dataset as a separate Gaussian distribution (left of Figure 3).

Distribution Modeling: Let $g_\theta(f(x_i))$ represent the final feature vector for a given input x_i. For each dataset d, we compute the mean and covariance in the projected space, assuming the distribution of the dataset follows a Gaussian distribution in the representation space:

$$\mu_d = \frac{1}{N_d}\sum_{i\in\mathcal{D}_d} z_i, \tag{4}$$

$$\Sigma_d = \frac{1}{N_d - 1}\sum_{i\in\mathcal{D}_d}(z_i - \mu_d)(z_i - \mu_d)^T \tag{5}$$

For datasets with limited samples, we employ Ledoit-Wolf shrinkage estimation (Ledoit and Wolf, 2004) to ensure numerical stability:

$$\hat{\Sigma} = (1-\lambda)\Sigma_{\text{sample}} + \lambda\frac{\text{tr}(\Sigma_{\text{sample}})}{d} I_d \tag{6}$$

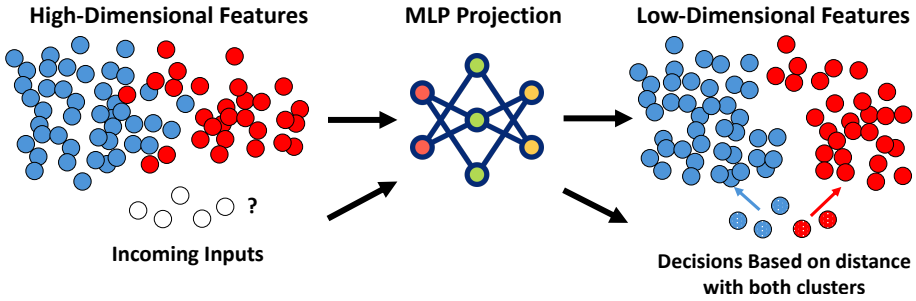


Figure 2: Safety-Aware projection for dimension reduction (Section 3.3)

Contrastive Scoring measures the relative proximity to benign vs. malicious distributions:

$$s_{\text{MCD}}(x) = \min_{d \in \text{benign}} D_M(g_\theta(f(x)), \mu_d, \Sigma_d) - \min_{d \in \text{malicious}} D_M(g_\theta(f(x)), \mu_d, \Sigma_d) \quad (7)$$

where $D_M(z, \mu, \Sigma) = \sqrt{(z-\mu)^T \Sigma^{-1} (z-\mu)}$ is the Mahalanobis distance. Higher scores indicate a greater likelihood of being malicious, as they imply that the given sample is closer to the malicious clusters in the representation space.

3.5 K-nearest Contrastive Detection (KCD)

KCD takes a non-parametric approach, making no distributional assumptions and minimal hyperparameters while being robust and effective. After normalizing features to the unit sphere, it computes distances to the k-th nearest neighbors with benign and malicious datasets (right of Figure 3):

$$s_{\text{KCD}}(x) = \|z - z_{(k)}^{\text{benign}}\|_2 - \|z - z_{(k)}^{\text{malicious}}\|_2 \quad (8)$$

where $z = g_\theta(f(x))/\|g_\theta(f(x))\|_2$ and $z_{(k)}$ denote the k-th nearest neighbor. This requires no covariance estimation and does not assume the distribution to be Gaussian. The intuition behind it is that for benign samples, the k-th nearest neighbor to the benign dataset should be closer, and similar for malicious samples. The parameter k is set to 50 to exclude noise and outliers.

3.6 Unified Decision Framework

Both methods use the same decision rule:

$$\delta(x) = \mathbb{1}[s(x) > \theta] \quad (9)$$

We calibrate the threshold θ on a held-out validation split drawn from the training data to maximize a weighted combination of balanced accuracy and F1 score. This calibration uses only training data and does not access the test set, ensuring an unbiased evaluation.

Remark: By incorporating both benign and malicious examples, our contrastive scoring approximates the log-likelihood ratio $\log \frac{p(x|\text{malicious})}{p(x|\text{benign})}$ needed for optimal Bayes decision making. This addresses the fundamental limitation of traditional OOD methods that only model the benign distribution (Nian et al., 2025). Section E formalizes this connection and shows that the score is an empirical Neyman–Pearson statistic.

4 Motivational Experiments

To highlight limitations in current evaluation protocols, we implement two simple OOD detection methods grounded in our discussion in Section 3.4 and Section 3.5, and compare them against state-of-the-art jailbreak detection systems using the JailDAM evaluation setup (Nian et al., 2025).

Experimental Setup: Following the JailDAM protocol, we utilize 414 MM-Vet-v2 samples (80% of the dataset, benign only) and evaluate the remaining 20% of MM-Vet-v2, as well as jailbreak attacks from MM-SafetyBench, FigStep, and JailbreakV-28K. Critically, our OOD methods use *no outlier exposure*—they are trained exclusively on benign samples, making this particularly challenging.

To validate our approach against alternative methods, we evaluate our LLaVA hidden state representations and FLAVA (Singh et al., 2022) embeddings. FLAVA is designed for multimodal understanding tasks and produces unified representations for both text-only and image-text inputs.

Implementation: We implement two simple methods based on previous sections: (1) KNN-based detection using 3-nearest neighbor distances in feature space and (2) Mahalanobis distance-based detection with regularized covariance estimation. Both methods use normalized features

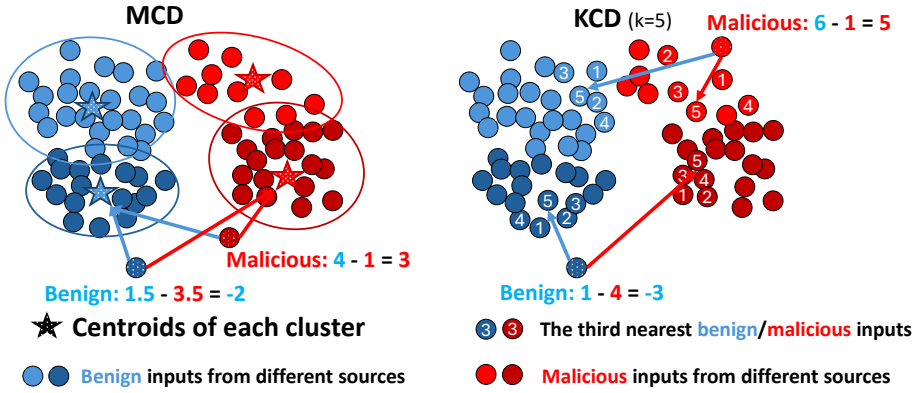


Figure 3: The proposed instantiations of RCS: MCD (left, Section 3.4) and KCD (right, Section 3.5).

Method	Model	Overall		MM-SBench		FigStep		JailbreakV	
		AR	AP	AR	AP	AR	AP	AR	AP
LLaVaGuard	Qwen	75.51	84.12	74.27	87.29	83.60	72.31	84.26	85.89
VLGuard	LLaVA	60.96	67.82	61.06	80.20	61.06	38.17	60.72	64.74
HiddenDetect	LLaVA	80.50	80.56	82.69	93.53	57.73	32.38	83.30	87.70
GradSafe	LLaVA	85.13	81.66	85.14	87.52	68.04	23.70	90.82	88.16
JailDAM*	CLIP	**95.50**	**95.30**	**91.26**	**98.04**	**96.08**	**96.16**	**94.65**	**94.64**
KNN-OOD	LLaVA	96.38	98.19	95.23	95.91	98.09	91.60	98.33	95.88
Mahal-OOD	LLaVA	**99.36**	**99.69**	**99.18**	**99.32**	**99.55**	**97.90**	**99.69**	**99.11**
KNN-OOD*	FLAVA	88.67	95.00	85.47	90.65	90.81	66.80	95.00	86.49
Mahal-OOD*	FLAVA	97.70	98.89	97.01	97.62	98.20	92.53	99.06	97.25

Table 1: Performance comparison (AR = AUROC, AP = AUPRC, MM-SBench is short for MM-SafetyBench) on JailDAM evaluation setup. Our simple methods use only benign training data and achieve superior performance. **Bold** and underlined values indicate the best performance and second-best among baseline methods. Methods with * don't use any hidden states from the target models.

extracted from LLaVA layer 16 (middle layer) and FLAVA embeddings. We remove the contrastive scoring and directly use the kNN distance/Mahalanobis distance to the benign cluster(s) to determine prompt safety. We do not include the learned projection stage because the dataset is too small.

Results: Table 1 reveals that our simple OOD methods outperform sophisticated defense mechanisms across all evaluation scenarios, given the same dataset access assumption with JailDAM. For instance, the Mahalanobis distance detector using LLaVA features achieves near-perfect performance, substantially surpassing the best baseline method previously proposed.

To further probe the limitations of one-class detection, we extend the evaluation of JailDAM (Nian et al., 2025). While its autoencoder-based approach performs well in a simplified setting, its reliance on modeling only in-distribution benign data makes it brittle to unseen, yet safe, inputs. We tested this by introducing unseen benign datasets from different domains (e.g., Medical VQA dataset VQA-RAD (Lau et al., 2018)). As shown in Table 2, the performance degrades significantly when faced with this benign distribution shift. The model's precision plummets as it incorrectly flags unseen benign samples as malicious, a critical issue of over-rejection. This underscores the need for methods that can distinguish *malicious intent* from *mere distribution shift*, a core motivation for our contrastive framework (depicted in Figure 4). We provide a more detailed setup and analysis in Section C.1.

Table 2: JailDAM performance in a simplified vs. robust evaluation scenario. The introduction of unseen benign data causes a sharp drop in precision, indicating a high false positive rate.

Evaluation Scenario	AUROC	Precision	Recall
Simplified (Original Setup)	0.9126	0.9491	0.9762
Robust (w/ Unseen Benign)	0.7057	0.5692	0.9452

Implications: These results indicate two critical insights: (1) Simply evaluating on individual attack datasets and one benign dataset may not adequately capture the complexity of real-world deployment, and (2) The safety-relevant information in LVLM representations is highly discriminative, suggesting that more powerful methods utilizing these features are needed to drive meaningful progress in the field. To better understand the difficulty of distinguishing malicious datasets from benign ones, we conducted a PCA analysis of the embedding spaces across different scenarios (Figure 6 in Section C.3).

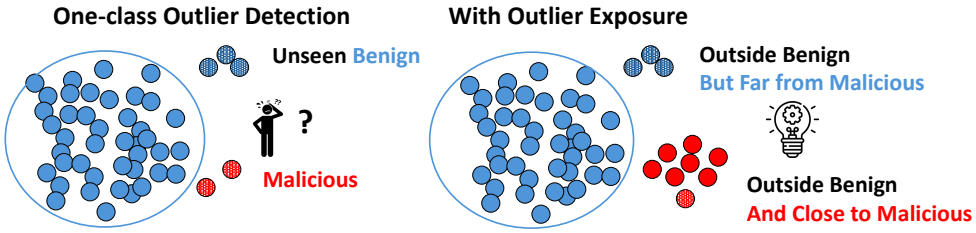


Figure 4: Comparison of one-class detection (Nian et al., 2025) and our proposed RCS with outlier exposure.

5 Experiments

5.1 Datasets and Experimental Scenarios

We construct a challenging benchmark with a balanced ratio of benign to malicious examples, drawing from a diverse mix of text-only and multimodal sources. The composition of our training and testing sets is provided in Table 6. A critical feature of our evaluation is the strict separation of attack types. For the JailbreakV-28K dataset, we train on two attack families but test exclusively on a third, held-out attack type. This protocol rigorously measures the ability to generalize to unseen attack strategies under distribution shift, preventing data leakage and ensuring a robust assessment.

FLAVA Baseline. To establish a strong, model-agnostic baseline, we train classifiers on embeddings from FLAVA (Singh et al., 2022), a model optimized for general multimodal understanding. This allows us to test a key hypothesis: does effective jailbreak detection require access to the target LVLM's internal reasoning states, which may contain specific *safety-related signals* from pre-existing safety alignment in LLMs? We extract the 768-dimensional embeddings and report the best-performing classifiers in our main results (Table 3). The unsatisfactory results, compared with KCD and MCD, prove that using target LVLM's internals is essential for detecting malicious behaviors.

5.2 Comprehensive Detection Performance

Following our principled layer selection, we evaluate our contrastive detection methods against supervised baselines on the optimal layers of `LLaVA-V1.6-Vicuna-7B` (Liu et al., 2023b) and `Qwen2.5-VL-7B` (Bai et al., 2025b). We also include results for `InternVL3-8B`[2] (Zhu et al., 2025b) in Section C.2. Our analysis focuses on layers selected by our principled layer selection method, which demonstrates consistently strong performance across all methods. We empirically demonstrate in Section D.5 that the reported range exhibits the strongest distinguishing performance for both Qwen and LLaVA. To ensure statistical reliability, we conduct 20 independent runs with different random seeds and report mean±std alongside the maximum values.

5.2.1 Experimental Protocol

We initialize and train layer-specific projections that reduce high-dimensional LVLM representations to 256 dimensions via multi-objective contrastive loss (Section 3.3). This projection, trained exclusively on training data, learns to cluster samples by dataset origin while maximizing benign–malicious separation. We pick and report the top-3 layers selected with our previously discussed strategy in Section 3.2. In Section D, we discuss the implementation in more detail, demonstrate the high effectiveness of the layer selection strategy, and illustrate its robustness when a noisier dataset, instead of SGXSTest, is adopted. For the KCD algorithm, we calculate a score based on the difference between the distances to the k (50 by default) nearest malicious and benign neighbors. MCD models each dataset as a Gaussian cluster with Ledoit-Wolf shrinkage (Ledoit and Wolf, 2004) for robust covariance estimation. Both methods operate on normalized ℓ_2 features. All training and testing are carried out on two NVIDIA RTX 4090 GPUs.

We evaluate existing state-of-the-art methods, including GradSafe (Xie et al., 2024), JailGuard (Zhang et al., 2023), HiddenDetect (Jiang et al., 2025a), and JailDAM (Nian et al., 2025), to serve as our baseline. Additionally, we include a variant of JailDAM, termed JailDAM-RCS, which applies our core contrastive idea to its framework and trains two parallel autoencoders—one on benign data and the other on malicious data—using the difference in reconstruction error as the detection score. We argue that it simultaneously measures the "outlierness" of both benign and malicious samples, which

[2]We use 'LLaVA', 'Qwen', and 'InternVL' for the rest of the paper for brevity, unless otherwise specified

Table 3: Detailed performance across different methods. We report mean±std (max/min for FPR) in percentages across 20 runs. The bold text presents the best performance among all layers and all methods.

Method	Layer	Classification				Separability	
		Accuracy(↑)	TPR(↑)	FPR(↓)	F1(↑)	AUROC(↑)	AUPRC(↑)
Target Model-Agnostic Methods							
K-Means	FLAVA	62.3	97.4	72.9	72.1	59.8	62.0
Logistic	FLAVA	59.8	37.6	18.0	48.3	61.7	66.8
JailDAM (Original)	CLIP	71.7	70.6	27.1	71.4	78.9	82.6
JailDAM-RCS	CLIP	84.5	93.4	24.4	85.8	91.5	90.0
LLaVA-v1.6-Vicuna-7B							
GradSafe	Critical Param.	66.5	96.9	64.9	74.1	75.4	79.4
HiddenDetect	16–29	81.6	79.9	16.8	81.2	90.1	90.0
JailGuard	Output	76.2	86.0	35.3	79.5	77.8	70.9
KCD	14	89.1±2.3 (94.3)	94.9±2.2 (97.2)	16.6±5.6 (2.3)	89.8±2.0 (94.4)	96.4±1.9 (96.2)	96.2±3.8 (98.6)
	15	89.4±2.3 (93.3)	95.3±2.7 (99.3)	16.5±6.1 (0.7)	90.0±1.9 (93.5)	96.9±2.1 (98.7)	96.3±4.1 (98.7)
	16	**92.0±2.1** (94.9)	94.1±3.6 (99.3)	**10.1±6.1** (1.7)	**92.2±1.8** (95.3)	97.7±0.9 (98.8)	97.2±1.2 (98.7)
MCD	14	88.3±2.2 (92.4)	95.5±1.9 (97.4)	18.9±5.8 (3.8)	89.1±1.7 (92.4)	97.4±0.3 (98.0)	97.5±0.3 (98.2)
	15	88.3±1.1 (91.1)	96.9±0.9 (99.0)	20.3±2.7 (14.4)	89.2±0.9 (91.5)	98.0±0.3 (98.5)	98.1±0.2 (98.5)
	16	91.0±2.3 (96.1)	**97.2±1.1** (98.7)	15.2±5.2 (2.8)	91.6±1.9 (96.1)	**98.6±0.1** (98.8)	**98.8±0.1** (98.7)
Qwen2.5-VL-7B							
GradSafe	Critical Param.	70.2	85.1	44.7	74.1	72.0	73.5
HiddenDetect	22–26	76.5	85.4	37.3	76.7	79.9	76.5
JailGuard	Output	56.1	64.1	51.9	59.4	44.0	45.4
KCD	20	87.3±1.7 (89.2)	97.2±2.0 (99.8)	22.5±4.5 (13.7)	88.5±1.2 (90.0)	96.1±3.3 (98.6)	94.8±3.5 (98.7)
	21	**89.2±2.7** (94.7)	96.1±2.7 (99.0)	**17.8±7.4** (2.1)	**90.0±2.2** (94.6)	96.3±3.5 (98.8)	93.6±4.2 (98.8)
	22	88.8±2.4 (94.5)	96.3±1.7 (99.0)	18.7±5.2 (7.9)	89.6±2.0 (94.6)	96.1±1.8 (98.6)	94.1±5.0 (98.5)
MCD	20	86.3±1.3 (88.2)	94.9±1.1 (99.4)	25.2±3.1 (8.3)	87.8±1.0 (89.2)	97.1±1.2 (98.5)	97.4±1.3 (98.7)
	21	86.6±1.5 (91.1)	98.1±1.0 (99.4)	24.8±3.3 (15.0)	88.0±1.2 (91.6)	97.7±0.9 (98.5)	**98.7±0.1** (98.9)
	22	87.0±2.3 (94.0)	**98.2±0.8** (99.3)	24.2±4.9 (8.8)	88.3±1.8 (94.2)	**98.1±0.6** (98.7)	98.3±0.6 (98.9)

is more powerful than solely using signals from benign samples. We defer the implementation details to Section F.

5.2.2 Results Analysis

We present the main results of LLaVA and Qwen in Table 3, which validate our central hypothesis that a contrastive approach is superior to one-class detection. Comparing the original JailDAM to our enhanced JailDAM-RCS, we see a dramatic performance leap of 16% in AUROC (78.9% to 91.5%). This demonstrates that the principle of modeling both benign and malicious distributions is highly effective on its own.

Second, our proposed methods, KCD and MCD, outperform even this strong contrastive baseline. MCD achieves a state-of-the-art 98.6% AUROC on LLaVA, detecting most of the malicious attempts and surpassing the enhanced JailDAM-RCS. Meanwhile, KCD achieves significantly lower false positive rates and superior F1 scores. This confirms that while the contrastive principle is crucial, its power is maximized when applied to the most discriminative signals—the internal geometric representations identified by our principled layer selection—rather than general-purpose embeddings like CLIP. Given that the hidden representation can be concurrently extracted with the inference process, the primary computational overhead of our method is associated with the learned projection from high-dimensional to low-dimensional space, as well as the computation of Mahalanobis distances or K-nearest neighbors. This overhead is minimal compared to the model's inference.

Note that for Qwen, JailGuard performs especially poorly. After investigation, we find that: 1) JailGuard works by perturbing the input, computing the divergence between responses, and marking samples with high response divergence as jailbreaks. For some attacks like VAE (Qi et al., 2024a), the model performs robustly across perturbations, rejecting the prompt most of the time; for others, some prompts repeatedly jailbreak the model. While the method fails for the latter case, it is actually safe in the former case. 2) JailGuard includes policies that rotate the pictures, which creates high false positives for the VizWiz dataset (Gurari et al., 2018) because the rotation can change the model's answer. We include examples for each case in the Appendix Section F.3.

5.2.3 Further Analysis

Ablation Studies. We conducted several ablation studies to validate our design choices in RCS. First, we show our choice of using the last-token hidden state, compared against mean pooling and last-5-

token aggregation, best captures the safety signals, whereas aggregation methods dilute discriminative performance (Section C.4.1). Furthermore, our experiments confirm that the learned, safety-aware projection significantly outperforms standard PCA and no dimensionality reduction (Section C.4.2). We also show that our methods are robust to hyperparameter choices, such as the clustering strategy in MCD (Section C.4.3) and the value of k in KCD (Section C.4.4). Finally, our sensitivity analysis (Section C.4.5) confirms that both dataset clustering (α) and safety separation (β) are essential in the loss term $\mathcal{L}$ in Section 3.3; setting either to zero results in a marked performance degradation. RCS also demonstrates robustness to weight variations, with optimal performance consistently observed at an $\alpha : \beta$ ratio of approximately 1:5, which we adopt for all main experiments.

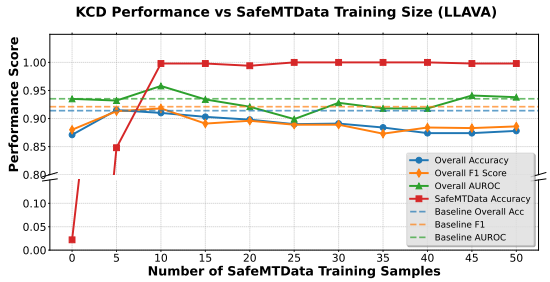


Figure 5: Detection performance of KCD vs. SafeMTData training size, tested over 5 runs on the optimal layer of LLaVA. Dashed lines indicate baseline performance without SafeMTData training and evaluation.

Adaptability to Multi-turn Jailbreaking Attacks. To evaluate the adaptability of our detection framework to unseen and challenging attack patterns, we conduct an additional experiment using SafeMTData (Ren et al., 2025), a dataset specifically designed for multi-turn jailbreaking scenarios with seemingly benign requests but increasingly malicious intentions, which represent a *near out-of-distribution* case. We vary the number of SafeMTData training samples from 0 to 50 while maintaining a fixed test set of 100 samples, with each configuration evaluated across 5 runs. The results of KCD and MCD for LLaVA (Figure 5 and Figure 14 in Section E.2) demonstrate remarkable adaptability: while both methods fail to detect when no training example is provided, with just 5-10 training samples, performance dramatically improves and reliable precision on original datasets is maintained. This highlights a key practical advantage of our approach: minimal exposure to new attack types enables rapid adaptation while preserving robustness against existing threats, making it particularly suitable for deployment scenarios where new attack patterns emerge continuously. See Section E.2 for additional results on Qwen and a sample-complexity analysis that explains the quick adaptation ability.

Computational Efficiency. To quantify the efficiency of our framework, we benchmarked the inference overhead of our detectors (including feature extraction, MLP projection, and scoring) against the standard forward pass of the host LVLM (LLaVA-V1.6-Vicuna-7B). As detailed in Table 4, the computational cost of our detector is negligible ($\leq 5.5\%$ relative overhead) compared to the inference time of the host LVLM. This comparison is conservative, as it does not account for the autoregressive generation phase; our method detects safety violations before the generation of the first token, potentially saving significant compute on rejected prompts. Notably, the K-NN search incurs less than 1% overhead and is highly parallelizable, ensuring scalability even with larger reference sets potentially needed for real-world robust deployment. Furthermore, the peak memory usage for the detector components is less than 0.015 GB, confirming that our method is highly efficient and lightweight.

Table 4: Inference efficiency benchmark.

Method	Component	Time (s)	Rel. Overhead
Baseline	LVLM Forward Pass	0.6383	-
KCD	MLP Projection	0.0220	3.4%
	K-NN Search	0.0040	0.6%
	Total KCD	**0.0260**	**~4.0%**
MCD	MLP Projection	0.0278	4.3%
	Mahalanobis Dist.	0.0077	1.2%
	Total MCD	**0.0355**	**~5.5%**

6 Conclusion

This work proposes Representational Contrastive Scoring (RCS), a general framework for jailbreak detection, with two novel instantiations that achieve state-of-the-art performance by applying simple statistical tests to the most geometrically discriminative internal layers of a model. Evaluated under a new, more realistic protocol that tests generalization to unseen attacks, our methods prove that effective, efficient, and generalizable safety does not require expensive retraining or complex external models, offering a practical path toward safer LVLM deployment.

7 Limitations

While our work presents a promising direction, we acknowledge several limitations.

Hyperparameter Tuning. First, our methods, though lightweight, still rely on hyperparameter selection (e.g., the value of k in KCD, the clustering strategy in MCD, and hyperparameters for the learned projection), which may require moderate tuning for optimal performance. Specifically, we observe relatively high variance during the training of our MLP projection network. We hypothesize that this stems from the over-parameterization of the MLP (projecting from about $d = 4096$ to $d = 256$ with around 1 million parameters), trained on a limited dataset of only 2,000 samples. Consequently, the optimization landscape is highly sensitive to random initialization across different runs. We believe that in real-world deployment, with much scaled-up training data, this variance would diminish, producing a more robust separation space; though large-scale data collection was outside our current scope. Future work could also explore more advanced model architectures and training recipes beyond supervised contrastive learning to enhance stability and performance.

Evaluation Scale. Our benchmark, while a significant step towards realism, serves as a proof-of-concept and is not exhaustive of the full spectrum of attacks and benign queries found in the wild. Therefore, a more comprehensive benchmark for jailbreak detection is still urgently needed. While real-world traffic is typically dominated by benign requests, our simplified benchmark utilizes a balanced 1:1 split to ensure statistical significance for AUROC and F1 metrics. In production, the decision threshold θ can be calibrated on a validation set to enforce a strict False Positive Rate (e.g., FPR $< 1\%$) regardless of the test set balance. Our high AUROC scores (approaching 0.99 for MCD) indicate that the method maintains a high True Positive Rate, even when the threshold is set strictly to suppress false alarms. We leave a more detailed discussion to future work.

White-box Dependency. Our requirement for internal representation of the protected model is an intentional design choice targeting safe model serving. This access enables detection at the "last input token position," effectively identifying jailbreaks *before* the model generates a response. This not only prevents the emission of toxic content but also saves the computational cost of full generation, which is often required by black-box output classifiers. Furthermore, as demonstrated in Section 3.2, these internal middle layers provide a significantly cleaner separation of benign and malicious inputs than surface-level embeddings.

Streaming Defense. We lastly note that while our method is compatible, we do not benchmark its *streaming defense* capabilities after the decoding stage starts. Recent streaming defense frameworks such as SCM (Li et al., 2025b) and Qwen3Guard-Stream (Zhao et al., 2025b) trade detection overhead for higher accuracy by real-time monitoring of model interactions and stop the generation when content is determined to be unsafe. We left the discussion of such modes to future work.

8 Broader Impacts and Ethical Considerations

The primary goal of this research is to enhance the safety and reliability of Large Vision-Language Models (LVLMs), a positive societal objective. By developing a lightweight and effective defense framework, Representational Contrastive Scoring (RCS), our work contributes to the responsible deployment of AI by mitigating the risks associated with jailbreak attacks. We hope this research empowers developers to build more robust safety guardrails, thereby preventing the generation of harmful content and increasing public trust in multimodal AI systems.

We acknowledge several ethical dimensions related to this work:

Defensive Nature. This research is strictly defensive. We do not introduce new attack vectors or datasets of harmful prompts. Instead, we focus on detecting and neutralizing existing threats, aiming to strengthen the security posture of LVLMs.

Data and Model Usage. All experiments were conducted using publicly available, open-weight models and established academic datasets. Our use of these artifacts is consistent with their licenses and intended research purposes, such as benchmarking model capabilities and safety evaluations. By design, the malicious datasets used for training and evaluation contain prompts intended to be harmful or offensive; this content is necessary for the explicit purpose of developing and testing a safety detector. The benign datasets were sourced

from established academic corpora that were previously vetted for public release. No new user data was collected, and no sensitive information is reproduced in this paper.

Potential for Adversarial Adaptation. While our method is a defense, any public research into safety mechanisms could potentially be studied by adversarial actors to devise more sophisticated attacks that attempt to circumvent it. We believe the benefit of sharing a strong, generalizable defense with the research community and developers outweighs this risk, as it fosters a more secure AI ecosystem overall.

False Positives and Over-Refusal. No detection system is perfect, and there remains a risk of false positives, where a benign prompt is incorrectly flagged as malicious. This could lead to unintended censorship or a degraded user experience. We show that our methods achieve a low false positive rate, and we stress that any real-world deployment of this technology should involve careful calibration of the detection threshold to balance safety with model utility.

Responsible Usage of AI in This Paper. This paper made limited use of generative AI tools (specifically, ChatGPT and Gemini) in accordance with the ACL Policy on Publication Ethics. The use was restricted to assistance purely with the language of the paper, such as paraphrasing and polishing for clarity and fluency. No generative AI tools were used to create, analyze, or interpret research content, and all substantive intellectual contributions were made solely by the authors.

Acknowledgment

We thank the reviewers for their valuable feedback. This work was partially supported by NSF (CNS-2154930, CNS-2403758), ARO (W911NF-24-1-0155), ONR (N000142412663), and Washington University.

References

Hervé Abdi and Lynne J Williams. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459.

Cem Anil, Esin Durmus, Nina Panickssery, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Meg Tong, Jesse Mu, Daniel Ford, Francesco Mosconi, Rajashree Agrawal, Rylan Schaeffer, Naomi Bashkansky, Samuel Svenningsen, Mike Lambert, Ansh Radhakrishnan, Carson Denison, Evan Hubinger, and 15 others. 2024. Many-shot jailbreaking. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *ArXiv preprint*, abs/2502.13923.

Valentyn Boreiko, Alexander Panfilov, Vaclav Voracek, Matthias Hein, and Jonas Geiping. 2024. An interpretable n-gram perplexity threat model for large language model jailbreaks. *ArXiv preprint*, abs/2410.16222.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. 2025. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 23–42. IEEE.

Jiawei Chen, Yang Yang, Chao Yu, Yu Tian, Zhi Cao, Xue Yang, Linghao Li, Hang Su, and Zhaoxia Yin. 2025a. Red teaming large reasoning models. *ArXiv preprint*, abs/2512.00412.

Taiye Chen, Zeming Wei, Ang Li, and Yisen Wang. 2025b. Scalable defense against in-the-wild jailbreaking attacks with safety context retrieval. *ArXiv preprint*, abs/2505.15753.

Yilun Chen, Ami Wiesel, and Alfred O Hero. 2011. Robust shrinkage estimation of high-dimensional covariance matrices. *IEEE Transactions on Signal Processing*, 59(9):4097–4107.

Jianfeng Chi, Ujjwal Karn, Hongyuan Zhan, Eric Smith, Javier Rando, Yiming Zhang, Kate Plawiak, Zacharie Delpierre Coudert, Kartikeya Upasani, and Mahesh Pasupuleti. 2024. Llama guard 3 vision: Safeguarding human-ai image understanding conversations. *ArXiv preprint*, abs/2411.10414.

Zachary Coalson, Jeonghyun Woo, Shiyang Chen, Yu Sun, Lishan Yang, Prashant Nair, Bo Fang, and Sanghyun Hong. 2024. Prisonbreak: Jailbreaking large language models with fewer than twenty-five targeted bit-flips. *ArXiv preprint*, abs/2412.07192.

Mahavir Dabas, Si Chen, Charles Fleming, Ming Jin, and Ruoxi Jia. 2025a. Just enough shifts: Mitigating over-refusal in aligned language models with targeted representation fine-tuning. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net.

Mahavir Dabas, Tran Huynh, Nikhil Reddy Billa, Jiachen T. Wang, Peng Gao, Charith Peris, Yao Ma, Rahul Gupta, Ming Jin, Prateek Mittal, and Ruoxi Jia. 2025b. Adversarial déjà vu: Jailbreak dictionary learning for stronger generalization to unseen attacks.

Taylor Denouden, Rick Salay, Krzysztof Czarnecki, Vahdat Abdelzad, Buu Phan, and Sachin Vernekar. 2018. Improving reconstruction autoencoder out-of-distribution detection with mahalanobis distance. *ArXiv preprint*, abs/1812.02765.

Zhihao Dou, Xin Hu, Haibo Yang, Zhuqing Liu, and Minghong Fang. 2024. Adversarial attacks to multi-modal models. In *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis, LAMPS 2024, Salt Lake City, UT, USA, October 14-18, 2024*, pages 35–46. ACM.

Xuefeng Du, Reshmi Ghosh, Robert Sim, Ahmed Salem, Vitor Carvalho, Emily Lawton, Yixuan Li, and Jack W Stokes. 2024. Vlmguard: Defending vlms against malicious prompts via unlabeled data. *ArXiv preprint*, abs/2410.00296.

Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. 2022. VOS: learning what you don't know by virtual outlier synthesis. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Kathleen C. Fraser, Hillary Dawkins, Isar Nejadgholi, and Svetlana Kiritchenko. 2025. Fine-tuning lowers safety and disrupts evaluation consistency. *ArXiv preprint*, abs/2506.17209.

Jiahui Gao, Renjie Pi, Tianyang Han, Han Wu, Lanqing Hong, Lingpeng Kong, Xin Jiang, and Zhenguo Li. 2024. Coca: Regaining safety-awareness of multimodal large language models with constitutional calibration. *ArXiv preprint*, abs/2409.11365.

Shaona Ghosh, Prasoon Varshney, Makesh Narsimhan Sreedhar, Aishwarya Padmakumar, Traian Rebedea, Jibin Rajan Varghese, and Christopher Parisien. 2025. AEGIS2.0: A diverse AI safety dataset and risks taxonomy for alignment of LLM guardrails. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5992–6026, Albuquerque, New Mexico. Association for Computational Linguistics.

Yichen Gong, Delong Ran, Xinlei He, Tianshuo Cong, Anyu Wang, and Xiaoyun Wang. 2025a. Safety misalignment against large language models. In *32nd Annual Network and Distributed System Security Symposium, NDSS 2025, San Diego, California, USA, February 24-28, 2025*. The Internet Society.

Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. 2025b. Figstep: Jailbreaking large vision-language models via typographic visual prompts. In *Thirty-Ninth AAAI Conference on Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 23951–23959. AAAI Press.

Weiyang Guo, Zesheng Shi, Zeen Zhu, Yuan Zhou, Min Zhang, and Jing Li. 2026. Backdoors in rlvr: Jailbreak backdoors in llms from verifiable reward. In *The 64th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Prannaya Gupta, Le Qi Yau, Hao Han Low, I-Shiang Lee, Hugo Maximus Lim, Yu Xin Teoh, Koh Jia Hng, Dar Win Liew, Rishabh Bhardwaj, Rajat Bhardwaj, and Soujanya Poria. 2024. WalledEval: A comprehensive safety evaluation toolkit for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 397–407, Miami, Florida, USA. Association for Computational Linguistics.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3608–3617. IEEE Computer Society.

Safayat Bin Hakim, Kanchon Gharami, Nahid Farhady Ghalaty, Shafika Showkat Moni, Shouhuai Xu, and Houbing Herbert Song. 2026. Jailbreaking llms: A survey of attacks, defenses and evaluation. *Authorea Preprints*.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. 2024. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural*

Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024.

Zeqing He, Zhibo Wang, Zhixuan Chu, Huiyu Xu, Wenhui Zhang, Qinglong Wang, and Rui Zheng. 2024. Jailbreaklens: Interpreting jailbreak mechanism in the lens of representation and circuit. *ArXiv preprint*, abs/2411.11114.

Zeqing He, Zhibo Wang, Huiyu Xu, and Kui Ren. 2025. Towards llm guardrails via sparse representation steering. *ArXiv preprint*, abs/2503.16851.

Lukas Helff, Felix Friedrich, Manuel Brack, Kristian Kersting, and Patrick Schramowski. 2024. Llavaguard: An open vlm-based framework for safeguarding vision datasets and models. *ArXiv preprint*, abs/2406.05113.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Dan Hendrycks, Mantas Mazeika, and Thomas G. Dietterich. 2019. Deep anomaly detection with outlier exposure. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Xiangdong Hu, Yangyang Jiang, Qin Hu, and Xiaojun Jia. 2026. Gambit: A gamified jailbreak framework for multimodal large language models. *ArXiv preprint*.

Rui Huang, Andrew Geng, and Yixuan Li. 2021. On the importance of gradients for detecting distributional shifts in the wild. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 677–689.

Joonhyun Jeong, Seyun Bae, Yeonsung Jung, Jaeryong Hwang, and Eunho Yang. 2025. Playing the fool: Jailbreaking llms and multimodal llms with out-of-distribution strategy. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29937–29946.

Jiabao Ji, Bairu Hou, Zhen Zhang, Guanhua Zhang, Wenqi Fan, Qing Li, Yang Zhang, Gaowen Liu, Sijia Liu, and Shiyu Chang. 2024. Advancing the robustness of large language models through self-denoised smoothing. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 246–257, Mexico City, Mexico. Association for Computational Linguistics.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Alex Qiu, Jiayi Zhou, Kaile Wang, Boxun Li, Sirui Han, Yike Guo, and Yaodong Yang. 2025. PKU-SafeRLHF: Towards multi-level safety alignment for LLMs with human preference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31983–32016, Vienna, Austria. Association for Computational Linguistics.

Fengqing Jiang, Zhangchen Xu, Luyao Niu, Zhen Xiang, Bhaskar Ramasubramanian, Bo Li, and Radha Poovendran. 2024. Artprompt: Ascii art-based jailbreak attacks against aligned llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15157–15173.

Yilei Jiang, Xinyan Gao, Tianshuo Peng, Yingshui Tan, Xiaoyong Zhu, Bo Zheng, and Xiangyu Yue. 2025a. Hiddendetect: Detecting jailbreak attacks against multimodal large language models via monitoring hidden states. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14880–14893.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. 2025b. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 25004–25014. Computer Vision Foundation / IEEE.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, and 196 others. 2025. Gemma 3 technical report.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. 2020. Why normalizing flows fail to detect out-of-distribution data. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Lingkai Kong, Haorui Wang, Wenhao Mu, Yuanqi Du, Yuchen Zhuang, Yifei Zhou, Yue Song, Rongzhi Zhang, Kai Wang, and Chao Zhang. 2024. Aligning large language models with representation editing: A control perspective. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. Openassistant conversations -

democratizing large language model alignment. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Aounon Kumar, Chirag Agarwal, Suraj Srinivas, Aaron Jiaxun Li, Soheil Feizi, and Himabindu Lakkaraju. 2023. Certifying llm safety against adversarial prompting. *ArXiv preprint*, abs/2309.02705.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.

Olivier Ledoit and Michael Wolf. 2004. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 7167–7177.

Taegyeong Lee, Jeonghwa Yoo, Hyoungseo Cho, Soo Yong Kim, and Yunho Maeng. 2025. QGuard:question-based zero-shot guard for multi-modal LLM safety. In *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*, pages 373–382, Vienna, Austria. Association for Computational Linguistics.

Jianwei Li and Jung-Eun Kim. 2025. Safety alignment can be not superficial with explicit safety signals. In *Forty-second International Conference on Machine Learning*.

Jiaqi Li, Qianshan Wei, Chuanyi Zhang, Guilin Qi, Miaozeng Du, Yongrui Chen, Sheng Bi, and Fan Liu. 2024. Single image unlearning: Efficient machine unlearning in multimodal large language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Jingyao Li, Pengguang Chen, Zexin He, Shaozuo Yu, Shu Liu, and Jiaya Jia. 2023a. Rethinking out-of-distribution (OOD) detection: Masked image modeling is all you need. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 11578–11589. IEEE.

Tianlong Li, Zhenghua Wang, Wenhao Liu, Muling Wu, Shihan Dou, Changze Lv, Xiaohua Wang, Xiaoqing Zheng, and Xuan-Jing Huang. 2025a. Revisiting jailbreaking for large language models: A representation engineering perspective. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3158–3178.

Xuan Li, Zhanke Zhou, Jianing Zhu, Jiangchao Yao, Tongliang Liu, and Bo Han. 2023b. Deepinception: Hypnotize large language model to be jailbreaker. *ArXiv preprint*, abs/2311.03191.

Yang Li, Qiang Sheng, Yehan Yang, Xueyao Zhang, and Juan Cao. 2025b. From judgment to interference: Early stopping llm harmful outputs via streaming content monitoring. *ArXiv preprint*, abs/2506.09996.

Yucen Lily Li, Daohan Lu, Polina Kirichenko, Shikai Qiu, Tim GJ Rudner, C Bayan Bruss, and Andrew Gordon Wilson. 2025c. Out-of-distribution detection methods answer the wrong questions. *ArXiv preprint*, abs/2507.01831.

Yucheng Li, Surin Ahn, Huiqiang Jiang, Amir H Abdi, Yuqing Yang, and Lili Qiu. 2025d. Securitylingua: Efficient defense of llm jailbreak attacks via security-aware prompt compression. *ArXiv preprint*, abs/2506.12707.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N Metaxas. 2025e. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. *ArXiv preprint*, abs/2502.03628.

Jiacheng Liang, Tanqiu Jiang, Yuhui Wang, Rongyi Zhu, Fenglong Ma, and Ting Wang. 2025. Autoran: Automated hijacking of safety reasoning in large reasoning models. *ArXiv preprint*, abs/2505.10846.

Shiyu Liang, Yixuan Li, and R. Srikant. 2018. Enhancing the reliability of out-of-distribution image detection in neural networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Yuping Lin, Pengfei He, Han Xu, Yue Xing, Makoto Yamada, Hui Liu, and Jiliang Tang. 2024. Towards understanding jailbreak attacks in llms: A representation space analysis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7067–7085.

Zihao Lin, Samyadeep Basu, Mohammad Beigi, Varun Manjunatha, Ryan A. Rossi, Zichao Wang, Yufan Zhou, Sriram Balasubramanian, Arman Zarei, Keivan Rezaei, Ying Shen, Barry Menglong Yao, Zhiyang Xu, Qin Liu, Yuxiang Zhang, Yan Sun, Shilong Liu, Li Shen, Hongxuan Li, and 2 others. 2025. A survey on mechanistic interpretability for multi-modal foundation models. *ArXiv preprint*, abs/2502.17516.

Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Yu Cheng, and Wei Hu. 2025a. A survey of attacks on large vision–language models: Resources, advances, and future trends. *IEEE Transactions on Neural Networks and Learning Systems*.

Han Liu, Yuhao Wu, Shixuan Zhai, Bo Yuan, and Ning Zhang. 2023a. Riatig: Reliable and imperceptible adversarial text-to-image generation with natural

prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20585–20594.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Qin Liu, Chao Shang, Ling Liu, Nikolaos Pappas, Jie Ma, Neha Anna John, Srikanth Doss, Lluis Marquez, Miguel Ballesteros, and Yassine Benajiba. 2024b. Unraveling and mitigating safety alignment degradation of vision-language models. *ArXiv preprint*, abs/2410.09047.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. 2024c. Reducing hallucinations in vision-language models via latent space steering. *ArXiv preprint*, abs/2410.15778.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. 2020. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. 2024d. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In *European Conference on Computer Vision*, pages 386–403. Springer.

Yi Liu, Junzhe Yu, Huijia Sun, Ling Shi, Gelei Deng, Yuqi Chen, and Yang Liu. 2024e. Efficient detection of toxic prompts in large language models. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, ASE '24, page 455–467, New York, NY, USA. Association for Computing Machinery.

Yue Liu, Xiaoxin He, Miao Xiong, Jinlan Fu, Shumin Deng, Yingwei Ma, Jiaheng Zhang, and Bryan Hooi. 2025b. Flipattack: Jailbreak llms via flipping. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net.

Yue Liu, Shengfang Zhai, Mingzhe Du, Yulin Chen, Tri Cao, Hongcheng Gao, Cheng Wang, Xinfeng Li, Kun Wang, Junfeng Fang, Jiaheng Zhang, and Bryan Hooi. 2025c. Guardreasoner-vl: Safeguarding vlms via reinforced reasoning. *ArXiv preprint*, abs/2505.11049.

Yule Liu, Zhen Sun, Xinlei He, and Xinyi Huang. 2024f. Quantized delta weight is safety keeper. *ArXiv preprint*.

Zhendong Liu, Yuanbi Nie, Yingshui Tan, Xiangyu Yue, Qiushi Cui, Chongjun Wang, Xiaoyong Zhu, and Bo Zheng. 2024g. Safety alignment for vision language models. *ArXiv preprint*, abs/2405.13581.

Weidi Luo, Siyuan Ma, Xiaogeng Liu, Xiaoyu Guo, and Chaowei Xiao. 2024. Jailbreakv-28k: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks. *CoRR*, abs/2404.03027.

Xiaoxu Ma, Xiangbo Zhang, and Zhenyu Weng. 2026. Stable and explainable personality trait evaluation in large language models with internal activations. *ArXiv preprint*, abs/2601.09833.

Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. 2022. Delving into out-of-distribution detection with vision-language representations. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Yichuan Mo, Yuji Wang, Zeming Wei, and Yisen Wang. 2024. Fight back against jailbreaking via prompt adversarial tuning. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Yi Nian, Shenzhe Zhu, Yuehan Qin, Li Li, Ziyi Wang, Chaowei Xiao, and Yue Zhao. 2025. Jaildam: Jailbreak detection with adaptive memory for vision-language model. *ArXiv preprint*, abs/2504.03770.

Vladimir Pestov. 2013. Is the k-nn classifier in high dimensions affected by the curse of dimensionality? *Computers & Mathematics with Applications*, 65(10):1427–1437.

Anirudh Phukan, Divyansh Divyansh, Harshit Kumar Morj, Vaishnavi Vaishnavi, Apoorv Saxena, and Koustava Goswami. 2025. Beyond logit lens: Contextual embeddings for robust hallucination detection & grounding in vlms. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9661–9675, Albuquerque, New Mexico. Association for Computational Linguistics.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. 2024a. Visual adversarial examples jailbreak aligned large language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, February 20-27, 2024, Vancouver, Canada*, pages 21527–21536. AAAI Press.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. 2025. Safety alignment should be

made more than just a few tokens deep. In *The Thirteenth International Conference on Learning Representations*.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2024b. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Cheng Qian, Hainan Zhang, Lei Sha, and Zhiming Zheng. 2025. Hsf: Defending against jailbreak attacks with hidden state filtering. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2078–2087.

Maan Qraitem, Nazia Tasnim, Piotr Teterwak, Kate Saenko, and Bryan A Plummer. 2024. Vision-llms can fool themselves with self-generated typographic attacks. *ArXiv preprint*, abs/2402.00626.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A. DePristo, Joshua V. Dillon, and Balaji Lakshminarayanan. 2019. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 14680–14691.

Qibing Ren, Hao Li, Dongrui Liu, Zhanxu Xie, Xiaoya Lu, Yu Qiao, Lei Sha, Junchi Yan, Lizhuang Ma, and Jing Shao. 2025. LLMs know their vulnerabilities: Uncover safety gaps through natural distribution shifts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24763–24785, Vienna, Austria. Association for Computational Linguistics.

Alexander Robey, Eric Wong, Hamed Hassani, and George J Pappas. 2023. Smoothllm: Defending large language models against jailbreaking attacks. *ArXiv preprint*, abs/2310.03684.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400, Mexico City, Mexico. Association for Computational Linguistics.

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Mark Russinovich, Ahmed Salem, and Ronen Eldan. 2025. Great, now write an article about that: The crescendo multi-turn LLM jailbreak attack. In *34th USENIX Security Symposium, USENIX Security 2025, Seattle, WA, USA, August 13-15, 2025*, pages 2421–2440. USENIX Association.

Christian Schlarmann and Matthias Hein. 2023. On the adversarial robustness of multi-modal foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3677–3685.

Vikash Sehwag, Mung Chiang, and Prateek Mittal. 2021. SSD: A unified framework for self-supervised outlier detection. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2024. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 1671–1685.

Abhay Sheshadri, Aidan Ewart, Phillip Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, and Stephen Casper. 2025. Latent adversarial training improves robustness to persistent harmful behaviors in llms. *Trans. Mach. Learn. Res.*, 2025.

Zesheng Shi, Yucheng Zhou, Jing Li, Yuxin Jin, Yu Li, Daojing He, Fangming Liu, Saleh Alharbi, Jun Yu, and Min Zhang. 2025. Safety alignment via constrained knowledge unlearning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 25515–25529. Association for Computational Linguistics.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15638–15650.

Jingran Su, Jingfan Chen, Hongxin Li, Yuntao Chen, Li Qing, and Zhaoxiang Zhang. 2025. Activation steering decoding: Mitigating hallucination in large vision-language models through bidirectional hidden state intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12964–12974.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. 2022. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 20827–20840. PMLR.

Yanchuan Tang, Taowen Wang, Yuefei Chen, Boxuan Zhang, Qiang Guan, and Ruixiang Tang. 2026. Shifting uncertainty to critical moments: Towards reliable uncertainty quantification for vla model.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Engkarat Techapanurak, Masanori Suganuma, and Takayuki Okatani. 2020. Hyperparameter-free out-of-distribution detection using cosine similarity. In *Proceedings of the Asian conference on computer vision*.

Roman Vershynin. 2018. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Cheng Wang, Yue Liu, Baolong Li, Duzhen Zhang, Zhongzhi Li, and Junfeng Fang. 2025a. Safety in large reasoning models: A survey. *ArXiv preprint*, abs/2504.17704.

Cheng Wang, Zeming Wei, Qin Liu, and Muhao Chen. 2025b. False sense of security: Why probing-based malicious input detection fails to generalize. *ArXiv preprint*, abs/2509.03888.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021. MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2140–2151, Online. Association for Computational Linguistics.

Xi Wang, Songlei Jian, Shasha Li, Xiaopeng Li, Zhaoye Li, Bin Ji, Baosheng Wang, and Jie Yu. 2026. Jpu: Bridging jailbreak defense and unlearning via on-policy path rectification.

Yanbo Wang, Jiyang Guan, Jian Liang, and Ran He. 2025c. Do we really need curated malicious data for safety alignment in multi-modal large language models? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19879–19889, Los Alamitos, CA, USA. IEEE Computer Society.

Yanting Wang, Hongye Fu, Wei Zou, and Jinyuan Jia. 2024a. Mmcert: Provable defense against adversarial attacks to multi-modal models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 24655–24664. IEEE.

Yu Wang, Xiaofei Zhou, Yichen Wang, Geyuan Zhang, and Tianxing He. 2024b. Jailbreak large vision-language models through multi-modal linkage. *ArXiv preprint*, abs/2412.00473.

Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. 2024. Assessing the brittleness of safety alignment via pruning and low-rank modifications. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Fenghua Weng, Jian Lou, Jun Feng, Minlie Huang, and Wenjie Wang. 2025a. Adversary-aware dpo: Enhancing safety alignment in vision language models via adversarial training. *ArXiv preprint*, abs/2502.11455.

Fenghua Weng, Yue Xu, Chengyan Fu, and Wenjie Wang. 2025b. Mmj-bench: A comprehensive study on jailbreak attacks and defenses for vision language models. In *Thirty-Ninth AAAI Conference on Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 27689–27697. AAAI Press.

Junfei Wu, Yue Ding, Guofan Liu, Tianze Xia, Ziyue Huang, Dianbo Sui, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2025. Sharp: Steering hallucination in lvlms via representation engineering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14357–14372.

Xinwei Wu, Weilong Dong, Shaoyang Xu, and Deyi Xiong. 2024. Mitigating privacy seesaw in large language models: Augmented privacy neuron editing via activation patching. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5319–5332.

Sophie Xhonneux, Alessandro Sordoni, Stephan Günnemann, Gauthier Gidel, and Leo Schwinn. 2024. Efficient adversarial training in llms with continuous attacks. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, Ruoxi Jia, Bo Li, Kai Li, Danqi Chen, Peter Henderson, and Prateek Mittal. 2025. SORRY-bench: Systematically evaluating large language model safety refusal. In *The Thirteenth International Conference on Learning Representations*.

Yueqi Xie, Minghong Fang, Renjie Pi, and Neil Gong. 2024. Gradsafe: Detecting jailbreak prompts for llms

via safety-critical gradient analysis. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 507–518, Bangkok, Thailand. Association for Computational Linguistics.

Xiqiao Xiong, Ouxiang Li, Zhuo Liu, Moxin Li, Wentao Shi, Fengbin Zhu, Qifan Wang, and Fuli Feng. 2026. Trojail: Trajectory-level optimization for multi-turn large language model jailbreaks with process rewards. *ArXiv preprint*.

Yue Xu, Xiuyuan Qi, Zhan Qin, and Wenjie Wang. 2024a. Cross-modality information check for detecting jailbreaking in multimodal large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13715–13726, Miami, Florida, USA. Association for Computational Linguistics.

Zhihao Xu, Ruixuan Huang, Changyu Chen, and Xiting Wang. 2024b. Uncovering safety risks of large language models through concept activation vector. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Chao Xue, Yao Wang, Mengqiao Liu, Di Liang, Xingsheng Han, Peiyang Liu, Xianjie Wu, Chenyao Lu, Lei Jiang, Yu Lu, Haibo Shi, Shuang Liang, Minlong Peng, and Flora D. Salim. 2026a. Reason only when needed: Efficient generative reward modeling via model-internal uncertainty. *ArXiv preprint*, abs/2604.10072.

Chao Xue, Yao Wang, Mengqiao Liu, Di Liang, Xingsheng Han, Peiyang Liu, Xianjie Wu, Chenyao Lu, Lei Jiang, Yu Lu, Haibo Shi, Shuang Liang, Minlong Peng, and Flora D. Salim. 2026b. Why supervised fine-tuning fails to learn: A systematic study of incomplete learning in large language models. *ArXiv preprint*, abs/2604.10079.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. 2024. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662.

Jirui Yang, Hengqi Guo, Zhihui Lu, Yi Zhao, Yuansen Zhang, Shijing Hu, Qiang Duan, Yinggui Wang, and Tao Wei. 2025. Prefix probing: Lightweight harmful content detection for large language models. *ArXiv preprint*, abs/2512.16650.

Mang Ye, Xuankun Rong, Wenke Huang, Bo Du, Nenghai Yu, and Dacheng Tao. 2025. A survey of safety on large vision-language models: Attacks, defenses and evaluations. *ArXiv preprint*, abs/2502.14881.

Jingwei Yi, Rui Ye, Qisi Chen, Bin Zhu, Siheng Chen, Defu Lian, Guangzhong Sun, Xing Xie, and Fangzhao Wu. 2024. On the vulnerability of safety alignment in open-access llms. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9236–9260, Bangkok, Thailand. Association for Computational Linguistics.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Fred Zhang and Neel Nanda. 2024. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Xiaocheng Zhang, Xi Wang, Yifei Lu, Jianing Wang, Zhuangzhuang Ye, Mengjiao Bao, Peng Yan, and Xiaohong Su. 2025a. Trendfact: A benchmark for explainable hotspot perception in fact-checking with natural language explanation. *ArXiv preprint*.

Xiaoyu Zhang, Cen Zhang, Tianlin Li, Yihao Huang, Xiaojun Jia, Ming Hu, Jie Zhang, Yang Liu, Shiqing Ma, and Chao Shen. 2023. Jailguard: A universal detection framework for llm prompt-based attacks. *ArXiv preprint*, abs/2312.10766.

Yihao Zhang, Zeming Wei, Jun Sun, and Meng Sun. 2024. Adversarial representation engineering: A general model editing framework for large language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Yongting Zhang, Lu Chen, Guodong Zheng, Yifeng Gao, Rui Zheng, Jinlan Fu, Zhenfei Yin, Senjie Jin, Yu Qiao, Xuanjing Huang, Feng Zhao, Tao Gui, and Jing Shao. 2025b. SPA-VL: A comprehensive safety preference alignment dataset for vision language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 19867–19878. Computer Vision Foundation / IEEE.

Chongwen Zhao, Zhihao Dou, and Kaizhu Huang. 2025a. Defending against jailbreak through early exit generation of large language models. In *International Conference on Neural Information Processing*, pages 532–546. Springer.

Haiquan Zhao, Chenhan Yuan, Fei Huang, Xiaomeng Hu, Yichang Zhang, An Yang, Bowen Yu, Dayiheng Liu, Jingren Zhou, Junyang Lin, Baosong Yang, Chen Cheng, Jialong Tang, Jiandong Jiang, Jianwei Zhang, Jijie Xu, Ming Yan, Minmin Sun, Pei Zhang, and 24 others. 2025b. Qwen3guard technical report. *ArXiv preprint*, abs/2510.14276.

Shiji Zhao, Ranjie Duan, Fengxiang Wang, Chi Chen, Caixin Kang, Shouwei Ruan, Jialing Tao, YueFeng Chen, Hui Xue, and Xingxing Wei. 2025c. Jailbreaking multimodal large language models via shuffle inconsistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2045–2054.

Wei Zhao, Zhe Li, Yige Li, Ye Zhang, and Jun Sun. 2024. Defending large language models against jailbreak attacks via layer-specific editing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5094–5109, Miami, Florida, USA. Association for Computational Linguistics.

Yiran Zhao, Wenxuan Zhang, Yuxi Xie, Anirudh Goyal, Kenji Kawaguchi, and Michael Shieh. 2025d. Understanding and enhancing safety mechanisms of LLMs via safety-specific neuron. In *The Thirteenth International Conference on Learning Representations*.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. 2023. On evaluating adversarial robustness of large vision-language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Hao Zheng, Zirui Pang, Ling li, Zhijie Deng, Yuhan Pu, Zhaowei Zhu, Xiaobo Xia, and Jiaheng Wei. 2025. Offside: Benchmarking unlearning misinformation in multimodal large language models.

Xiaosen Zheng, Tianyu Pang, Chao Du, Qian Liu, Jing Jiang, and Min Lin. 2024. Improved few-shot jailbreaking can circumvent aligned language models and their defenses. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Andy Zhou, Bo Li, and Haohan Wang. 2024a. Robust prompt optimization for defending language models against jailbreaking attacks. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Yi Zhou, Wenpeng Xing, Dezhang Kong, Changting Lin, and Meng Han. 2025. Neurel-attack: Neuron relearning for safety disalignment in large language models. *ArXiv preprint*, abs/2504.21053.

Yibo Zhou. 2022. Rethinking reconstruction autoencoder-based out-of-distribution detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 7369–7377. IEEE.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, and Yongbin Li. 2024b. How alignment and jailbreak work: Explain LLM safety through intermediate hidden states. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, Findings of ACL, pages 2461–2488. Association for Computational Linguistics.

Boyu Zhu, Xiaofei Wen, Wenjie Jacky Mo, Tinghui Zhu, Yanan Xie, Peng Qi, and Muhao Chen. 2025a. Omniguard: Unified omni-modal guardrails with deliberate reasoning. *ArXiv preprint*, abs/2512.02306.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025b. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *ArXiv preprint*, abs/2504.10479.

Yongshuo Zong, Ondrej Bohdal, Tingyang Yu, Yongxin Yang, and Timothy M. Hospedales. 2024. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023a. Representation engineering: A top-down approach to ai transparency. *ArXiv preprint*, abs/2310.01405.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023b. Universal and transferable adversarial attacks on aligned language models. *ArXiv preprint*, abs/2307.15043.

A Detailed Related Work

In this section, we provide a detailed literature review accompanied by our proposed taxonomy. It is important to note that neither the categorization of attacks nor that of defenses is unique, and the present review does not aim to be exhaustive. For more comprehensive systematizations of the field, readers are referred to existing survey articles, such as (Liu et al., 2025a; Ye et al., 2025; Hakim et al., 2026).

Jailbreak Attacks. Jailbreak attacks against LLMs have evolved from simple manual prompt crafting to sophisticated automated techniques, including optimization-based attacks (Zou et al., 2023b; Chao et al., 2025; Zhao et al., 2023), role-playing attacks (Li et al., 2023b), token manipulation attacks (Jiang et al., 2024; Liu et al., 2025b), in-context learning (Anil et al., 2024; Zheng et al., 2024), multi-turn conversation exploits (Chao et al., 2025; Russinovich et al., 2025; Ren et al., 2025), reasoning hijacking (Liang et al., 2025; Chen et al., 2025a; Wang et al., 2025a), backdoors (Guo et al., 2026), and hardware fault injection (Coalson et al., 2024). Sorry-bench (Xie et al., 2025) evaluates the refusal behaviors of LLMs against unsafe prompts

under a fine-grained taxonomy and diverse strategies, such as linguistic characteristics and the formatting of prompts.

Multimodal jailbreaks introduce additional complexity by exploiting the vision-language interface through adversarial or out-of-distribution images (Qi et al., 2024a; Jeong et al., 2025; Dou et al., 2024), prompt manipulation (Zhao et al., 2025c; Qraitem et al., 2024), and cross-modal prompt injection (Gong et al., 2025b; Wang et al., 2024b). Recent work further targets the model's reasoning process itself, for example by embedding harmful intent within gamified, puzzle-like tasks that induce cognitive overload and goal-driven reasoning, thereby reducing safety awareness and increasing attack success rates (Hu et al., 2026). There have been multiple works that attempt to explore a systematic and comprehensive evaluation of the robustness of multimodal large language models. For example, JailbreakV-28K (Luo et al., 2024), MMJ-Bench (Weng et al., 2025b), and MM-SafetyBench (Liu et al., 2024d) provide large-scale datasets of LLM-transfer attacks and query-relevant image attacks. They highlight the increasing sophistication and diversity of attack strategies, motivating the need for robust, generalizable detection methods and evaluation frameworks.

Jailbreak Defense. Defense mechanisms against jailbreak attacks have evolved across multiple levels of the model pipeline. **Input-level defenses** filter or transform potentially malicious prompts through statistical validation (e.g. N-gram Perplexity (Boreiko et al., 2024)), security-aware compression (SecurityLingua (Li et al., 2025d)), embedding-level detection (Liu et al., 2024e), or adversarial perturbation mitigation (Robey et al., 2023; Ji et al., 2024; Xu et al., 2024a). More recently, OMNIGUARD (Zhu et al., 2025a) attempts to train a guardrail LLM on inputs of any combination of input modalities. The primary advantage of input-level defenses is that they enable decision-making without requiring the model to produce a full output sequence, thereby potentially reducing the computational cost associated with decoding and preventing harmful responses from reaching users. However, OMNIGUARD requires reasoning on the multimodal inputs, which may significantly introduce latency for long contexts. In a similar vein, **prompt engineering** defenses provide deployment-time flexibility, such as QGuard (Lee et al., 2025), prefix probing (Yang et al., 2025), prompt optimization (Zhou et al., 2024a; Mo et al., 2024), token erasure (Kumar et al., 2023), and dynamic safety context retrieval (Chen et al., 2025b). **Output-level classifiers** usually employ specialized LLMs as safety guardians. For example, Qwen3guard (Zhao et al., 2025b), WildGuard (Han et al., 2024), and GuardReasoner (Liu et al., 2025c) achieve state-of-the-art performance in dual prompt/response classification. SCM (Li et al., 2025b) and Qwen3Guard-Stream (Zhao et al., 2025b) propose detecting harmful output in streaming and can abort generation when only part of the outputs has been generated. In contrast, our detector can reliably detect malicious intents before generation by leveraging hidden representations in the model. **Alignment-based** approaches fundamentally reshape model behavior, examples include Direct Preference Optimization (DPO) variants (Zhang et al., 2025b; Weng et al., 2025a)), adversarial tuning (Ghosh et al., 2025; Xhonneux et al., 2024; Sheshadri et al., 2025), safety-aware RLHF methods (Ji et al., 2025), and safety unlearning (Zheng et al., 2025; Li et al., 2024; Shi et al., 2025; Wang et al., 2026). ACTOR (Dabas et al., 2025a) fine-tunes the model in the representation space to mitigate over-refusal behaviors; while their focus on representational analysis is similar to ours, we primarily investigate jailbreak detection that does not modify the model internals. Adversarial Déjà Vu (Dabas et al., 2025b) attempts to extract "skill set" from existing jailbreak prompts and generalize the alignment to unseen attacks with dictionary learning and explanation. While showing good generalization capabilities, their techniques include high costs in building the skill set, dictionary learning, and model alignment. Beyond alignment itself, some works explore the alignment-breaking phenomenon (safety misalignment) (Qi et al., 2024b; Fraser et al., 2025; Wei et al., 2024; Gong et al., 2025a). Liu et al. show that delta-weight quantization in fine-tuned LLMs can reduce alignment-breaking and backdoor risks while lowering serving overhead (Liu et al., 2024f). These efforts focus on text-only prompts, and the applicability to multimodal LLMs remains unknown. Recently, a line of work discusses how we can enable LVLMs to inherently possess the safety capabilities of the backbone LLMs (Liu et al., 2024b; Gao et al., 2024; Wang et al., 2025c). VLGuard (Zong et al., 2024) and LLaVAGuard (Helff et al., 2024) are the leading efforts that propose curated datasets and training recipes specifically for large vision language

models. While these defenses show promise, they primarily focus on specific attack vectors or modalities in isolation, whereas our work provides a unified detection framework that operates across diverse attack strategies and input modalities through distributional analysis.

Representation Engineering. Recent work has demonstrated that LLM intermediate representations encode rich semantic information about input intent and safety (Zou et al., 2023a; Arditi et al., 2024; Zhou et al., 2024b; Li et al., 2025a; Lin et al., 2024; Du et al., 2024; Xu et al., 2024b). Researchers have applied this principle to hallucination detection and mitigation (Liu et al., 2024c; Wu et al., 2025; Li et al., 2025e; Su et al., 2025), uncertainty quantification (Tang et al., 2026; Xue et al., 2026a), personality trait evaluation (Ma et al., 2026), fact tracking (Zhang et al., 2025a), and model editing (Zhang et al., 2024; Kong et al., 2024), among others. Activation patching (Wu et al., 2024; Zhang and Nanda, 2024) shows that specific layers or neurons correlate with harmful content generation, with several studies examining the interactions between alignment and safety-critical neurons (Wei et al., 2024; Zhou et al., 2025) or layers (Zhao et al., 2024, 2025d). While several works empirically demonstrate that benign and malicious prompts are separable in intermediate layer representations (Zhou et al., 2024b; He et al., 2024; Du et al., 2024; Qian et al., 2025), these studies are limited in scope—typically evaluating single attack patterns on small-scale datasets with minimal diversity in benign samples. Meanwhile, recent concurrent work (Wang et al., 2025b) suggests that probing-based classifiers may rely on superficial linguistic patterns and trigger words, potentially leading to a "false sense of security" when facing significant distribution shifts. Moreover, existing representation-based detection work predominantly focuses on text-only LLMs, overlooking the unique challenges posed by multimodal inputs in LVLMs. Our work builds on these insights by extracting features from safety-critical layers (cf. Section 3.2 and Section D) and modeling their distributional geometry for jailbreak detection; however, it critically extends prior efforts through: (i) comprehensive evaluation across diverse attack strategies and benign datasets spanning both text-only and multimodal inputs (cf. Section 5.1), (ii) an explicit focus on LVLMs, which must handle both modalities seamlessly, and (iii) deployment-oriented evaluation protocols that reflect realistic distribution shifts when encountering unseen datasets. We also show in Section 5.2.3 that our methods are highly adaptable to unseen, more challenging datasets (Ren et al., 2025).

Out-of-Distribution (OOD) Detection. OOD detection aims to identify test samples that differ significantly from training distributions (Yang et al., 2024). Modern approaches are categorized into four main types: **Classification-based methods** leverage model outputs for detection, including post-hoc approaches like maximum softmax probability (MSP) (Hendrycks and Gimpel, 2017), ODIN (Liang et al., 2018), and energy-based scores (Liu et al., 2020), as well as training-based methods with outlier exposure (Hendrycks et al., 2019) or virtual outlier synthesis (Du et al., 2022). **Distance-based methods** compute distances to ID prototypes using Mahalanobis distance (Lee et al., 2018; Sehwag et al., 2021), cosine similarity (Techapanurak et al., 2020), or non-parametric k-NN approaches (Sun et al., 2022). **Density-based methods** explicitly model ID distributions through Gaussian mixtures (Lee et al., 2018), normalizing flows (Kirichenko et al., 2020), or likelihood ratios (Ren et al., 2019), though they often struggle with high-dimensional spaces. **Reconstruction-based methods** exploit differences in reconstruction quality between ID and OOD samples (Denouden et al., 2018; Zhou, 2022; Li et al., 2023a). Recent work also explores gradient-based detection (Huang et al., 2021) and foundation model adaptations (Ming et al., 2022).

B Details of the Constructed Benchmark

We disclose the details of our constructed challenging benchmark in Table 6. The benchmark contains both benign and malicious, text-only and multimodal inputs, divided into the training and testing samples. For the ease of testing and interpretation of results, we deliberately make the ratio of benign and malicious samples 1:1.

C Additional Results

C.1 Detailed Analysis of JailDAM's Limitations

To understand the practical limitations of existing black-box detection methods, we conducted a detailed analysis of JailDAM (Nian et al., 2025), a state-of-the-art approach that uses an autoencoder to detect jailbreaks without requiring access to

Table 5: Detailed performance across different methods. We report mean±std (max/min for FPR) in percentages across 20 runs. The bold text presents the best performance among all layers and all methods.

Method	Layer	Classification				Separability	
		Accuracy(↑)	TPR(↑)	FPR(↓)	F1(↑)	AUROC(↑)	AUPRC(↑)
Target Model-Agnostic Methods							
K-Means	FLAVA	62.3	97.4	72.9	72.1	59.8	62.0
Logistic	FLAVA	59.8	37.6	18.0	48.3	61.7	66.8
JailDAM (Original)	CLIP	71.7	70.6	27.1	71.4	78.9	82.6
JailDAM-RCS	CLIP	84.5	93.4	24.4	85.8	91.5	90.0
InternVL3-8B							
GradSafe	Critical Param.	69.1	92.9	54.7	62.9	80.2	79.5
HiddenDetect	18–24	62.5	68.7	57.1	54.6	75.9	78.4
JailGuard	Output	69.2	60.0	21.6	72.6	76.2	77.5
KCD	20	87.7±0.9 (87.2)	97.0±1.6 (98.0)	21.6±5.6 (13.3)	88.6±3.1 (92.4)	93.0±3.6 (95.3)	92.9±3.4 (95.7)
	21	**89.5±1.7** (93.3)	97.5±1.4 (99.1)	**15.5±4.1** (10.7)	89.1±2.9 (94.5)	92.6±4.1 (96.4)	92.4±3.3 (96.3)
	22	88.4±1.7 (92.1)	**97.6±1.3** (98.7)	20.8±6.1 (9.2)	89.3±1.5 (92.3)	92.5±2.1 (95.8)	92.2±3.7 (95.7)
MCD	20	88.7±4.2 (90.3)	97.5±1.9 (99.6)	20.1±5.8 (6.8)	**89.4±0.7** (93.4)	95.0±3.2 (96.9)	94.0±3.7 (96.5)
	21	89.1±1.3 (91.1)	95.2±0.6 (98.0)	17.0±2.7 (10.4)	88.6±1.1 (91.2)	**96.2±0.6** (96.8)	**96.0±0.6** (96.7)
	22	88.0±2.5 (88.2)	96.0±1.0 (98.6)	20.0±5.2 (12.8)	88.3±1.2 (90.1)	96.0±1.0 (97.0)	95.9±1.1 (97.1)

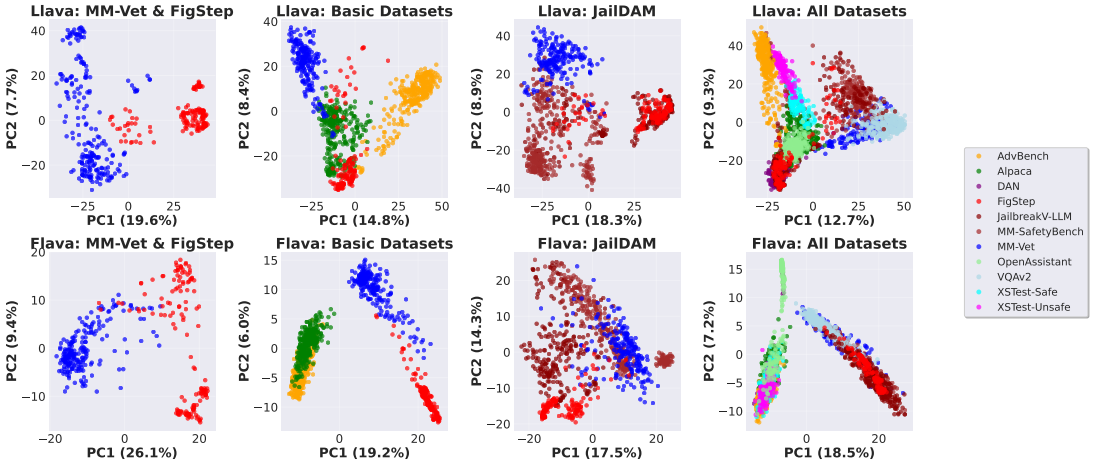
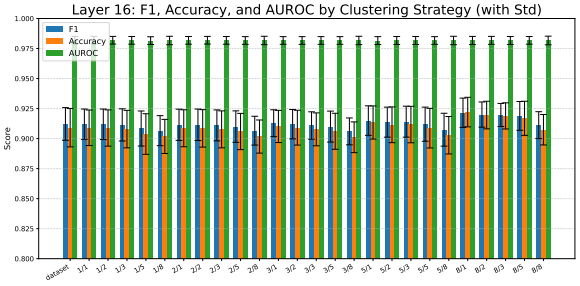
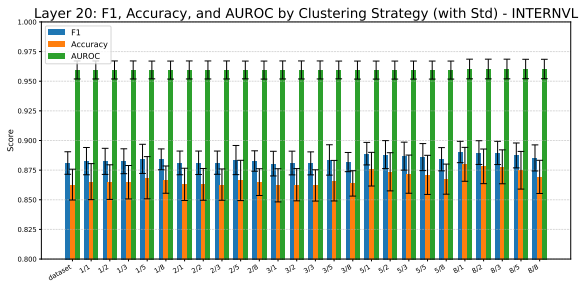


Figure 6: PCA visualization revealing the underlying separability structure across different evaluation scenarios. **Top row:** LLaVA layer 16 embeddings. **Bottom row:** FLAVA embeddings. **Columns (left to right):** (1) Binary case with MM-Vet vs JailbreakV-FigStep, (2) Basic datasets including text-only and multimodal samples, (3) JailDAM evaluation setup, (4) Full complexity with all available datasets. The JailDAM setup (third column) exhibits clear linear separability between benign (blue) and malicious (red/brown) clusters, explaining why simple OOD methods achieve near-perfect performance. In contrast, the full dataset scenario (fourth column) reveals substantial overlap and complex manifold structure, representing a more realistic and challenging evaluation setting.



(a) Clustering strategies for MCD on layer 16 (LLaVA).



(b) Clustering strategies for MCD on layer 20 (InternVL).

Figure 7: Performance comparison of clustering strategies for MCD on layer 16 (LLaVA) and layer 20 (InternVL). Different k_benign/k_malicious ratios are tested against the baseline dataset-based approach. The 8/1 configuration (8 benign clusters, 1 malicious cluster) achieves optimal performance in both cases. Error bars show standard deviation across multiple runs.

Table 6: Composition of the training and testing datasets. Our setup ensures a diverse mix of modalities and a strict separation of attack types for JailbreakV-28K.

Split	Class	Source	Modality	# Samples
Training	Benign	Alpaca (Taori et al., 2023)	Text	500
		MM-Vet (Yu et al., 2024)	MM*	218
		OpenAssistant (Köpf et al., 2023)	Text	282
	Malicious	AdvBench (Zou et al., 2023b)	Text	300
		DAN (Shen et al., 2024)	Text	150
		JailbreakV-28K* (Luo et al., 2024)	MM	550
Testing	Benign	XSTest (Röttger et al., 2024)	Text	250
		FigTxt (Jiang et al., 2025a)	Text	300
		VizWiz (Gurari et al., 2018)	MM	450
	Malicious	XSTest (Röttger et al., 2024)	Text	200
		FigTxt (Jiang et al., 2025a)	Text	350
		VAE (Qi et al., 2024a)	MM	200
		JailbreakV-28K† (Luo et al., 2024)	MM	150

* "MM" here means the prompt samples in the dataset are multimodal.

* Training split uses "LLM Transfer" and "Query-Related" attack types.

† Testing split uses the held-out "FigStep" attack type to measure generalization.

harmful training data. Our investigation reveals a fundamental vulnerability in its one-class detection design: an inability to distinguish between malicious inputs and benign inputs under distribution shift.

C.1.1 The Root Cause: Conflating Distribution Shift with Malicious Intent

JailDAM's core mechanism is to train an autoencoder exclusively on in-distribution (ID) benign data (e.g., the MM-Vet dataset (Yu et al., 2024)). The model learns to reconstruct the attention features derived from these safe inputs effectively. The guiding assumption is that malicious jailbreak attempts will produce feature representations that the autoencoder cannot reconstruct well, resulting in a high reconstruction error score that flags the input as anomalous (and therefore harmful).

While sound in a closed-world setting, this logic breaks down in open-world scenarios where the detector encounters benign data from domains not seen during training. An OOD benign sample, such as a medical image from the VQA-RAD dataset (Lau et al., 2018), is, by definition, distributionally different from the general-domain images in MM-Vet. Consequently, the autoencoder, having never learned to represent features from the medical domain, assigns a high reconstruction error to these samples. It conflates benign **distribution shift** with malicious **intent**.

This phenomenon is vividly illustrated in our experiments. Figure 8 shows the score distribution in the simplified, original evaluation setting. Here, the benign validation set (orange) is well-separated from the unsafe set (blue). However, Figure 9 shows the distribution in our more robust setting. The scores of the unseen benign VQA-RAD dataset (green) almost completely overlap with those of the unsafe malicious dataset (orange), making them nearly indistinguishable on the basis of reconstruction error alone. The text-only benign instructions (blue), being simpler, maintain lower error scores; however, the VQA-RAD distribution demonstrates the critical failure point.

C.1.2 Experimental Results and Discussion

We formalized this analysis in two experiments.

Experiment 1: Simplified Scenario (Original JailDAM Setup) We trained the JailDAM autoencoder on 80% of the benign MM-Vet dataset. We then tested its ability to distinguish between the remaining 20% of MM-Vet (as a validation set) and the malicious MM-SafetyBench (Liu et al., 2024d) dataset. The results confirm the high performance reported in the original paper, as shown in Table 7.

Experiment 2: Robust Scenario with OOD Benign Data Using the exact same trained model, we expanded the test set to include two unseen benign datasets: VQA-RAD (a medical VQA dataset (Lau et al., 2018)) and a set of text-only instructions

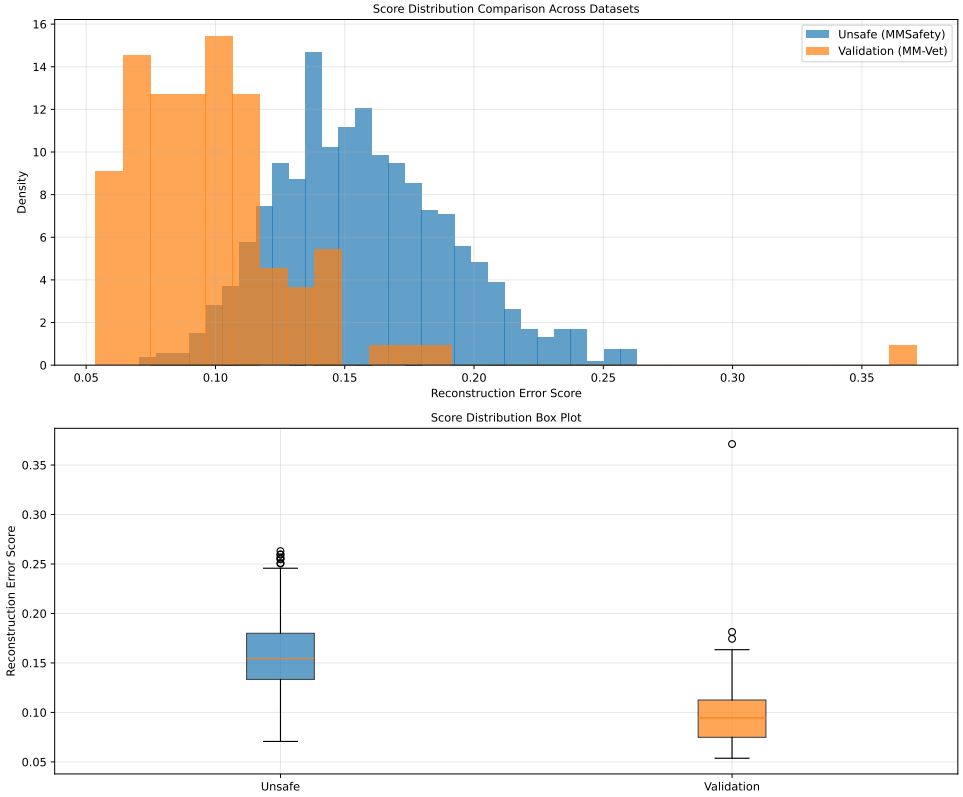


Figure 8: Score distribution of the original JailDAM method in a simplified evaluation setting. The plot shows that the reconstruction error scores for the benign set are concentrated at lower values, while scores for the malicious set are higher, indicating clear geometric separability in this controlled scenario.

Table 7: JailDAM Performance in Simplified Scenario.

Metric	Value
AUROC	0.9126
AUPRC	0.9804
F1 Score	0.9624
Precision	0.9491
Recall	0.9762

(from Alpaca (Taori et al., 2023)). This tests the model's robustness to both domain shift (medical images) and modality shift (text-only).

Table 8: JailDAM Performance in Robust Scenario.

Metric	Value
AUROC	0.7057
AUPRC	0.6072
F1 Score	0.7105
Precision	0.5692
Recall	0.9452

The performance collapses dramatically, as shown in Table 8. The AUROC drops by over 20 points. Critically, while recall remains high (94.5%), precision falls to 56.9%. This indicates severe **over-rejection**: the model correctly identifies most malicious inputs, but at the cost of incorrectly flagging a vast number of legitimate, unseen benign inputs as harmful. A per-dataset breakdown reveals the cause: the mean reconstruction score for the unseen benign VQA-RAD data (0.1660) is substantially higher than that of the ID benign data (0.0990) and is statistically much closer to the mean score of the malicious data (0.1576).

C.1.3 Conclusion

This analysis demonstrates that the original one-class autoencoder method of JailDAM, while effective in a controlled environment, is fundamentally unsuited for diverse applications where it may encounter benign inputs from unseen distributions. However, this does not invalidate the underlying principle of using reconstruction error as a discriminative signal.

In fact, our own evaluation in Table 3 and Table 5 shows that when the autoencoder architecture is integrated into our Representational Contrastive Scor-

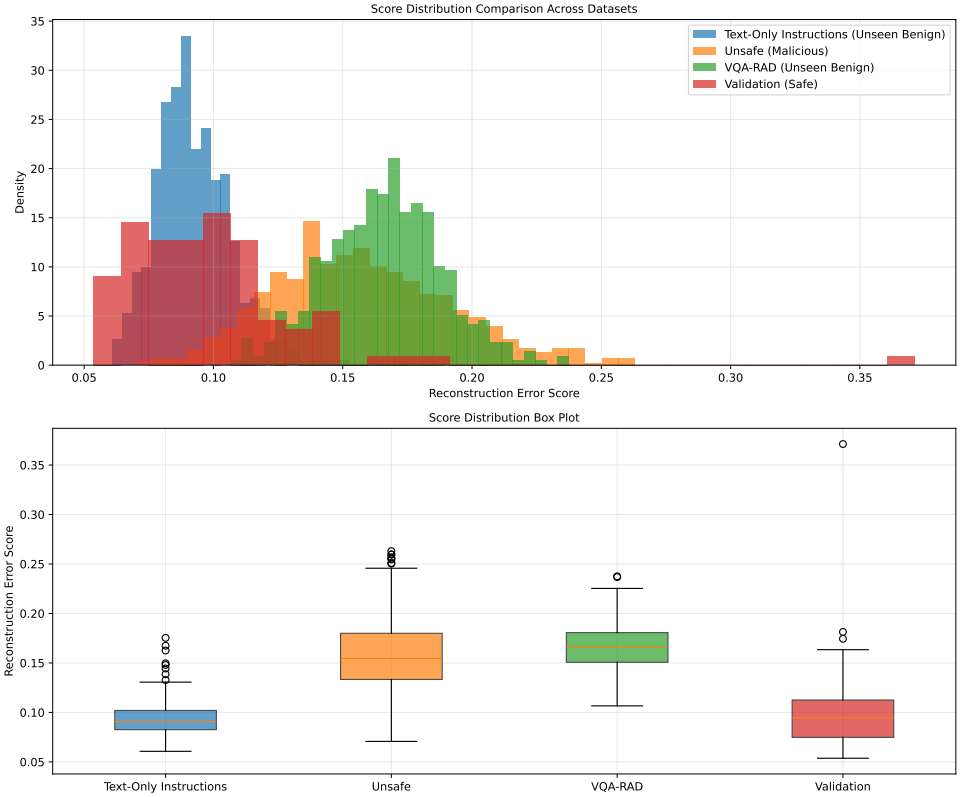


Figure 9: Score distribution of JailDAM in our robust evaluation setting, which introduces out-of-distribution (OOD) benign data. The reconstruction errors for the unseen benign VQA-RAD dataset significantly overlap with the malicious Unsafe dataset.

ing framework (JailDAM-RCS), its performance is significantly enhanced, with AUROC jumping from 78.9% to 91.5% on our challenging benchmark. This strongly suggests that the limitation lies not in the reconstruction-based approach itself, but in its one-class application. By training separate models for benign and malicious distributions and using a contrastive score (more details in Section F.1), the system learns to differentiate true malicious intent from mere distribution shift. Our work further advances this principle by applying a similar contrastive logic not to external embeddings, but to the richer and more discriminative internal representations of the LVLM itself.

C.2 Results for InternVL3

Apart from the results in Table 3, we also provide a comparison of the results for InternVL3-8B (Zhu et al., 2025b) in Table 5. We keep the target model-agnostic methods for reference.

C.3 PCA Visualization of Dataset Separability

To better understand the difficulty of distinguishing malicious datasets from benign ones, we conducted a PCA analysis of the embedding spaces across different evaluation scenarios.

In the first column, we include only one malicious and one benign dataset, corresponding to the last three columns of Table 1. In the second column, we include two benign and two malicious datasets; in each category, one is text-only and the other is multimodal. We note that the text-only benign dataset, Alpaca, has significant overlap with two malicious datasets (FigStep and AdvBench), respectively. This brings up an issue ignored by JailDAM: the inputs to the LVLMs can be unimodal or multimodal, and distinguishing between multimodal benign datasets and attacks only captures a part of real-world cases.

The visualization also reveals that datasets in the JailDAM setup (third column) exhibit clustering and separability between benign MM-Vet samples (blue) and malicious datasets (red/brown clusters), particularly evident in both LLaVA layer 16. In contrast, when we examine the full dataset complexity in our setting (rightmost column, covered in the next section), the embedding space reveals substantial overlap between benign and malicious samples, complex manifold structures, and ambigu-

ous boundary regions that would challenge any detection method. This stark difference in complexity validates our argument that current evaluation protocols may be insufficient for assessing real-world robustness.

C.4 Ablation Studies

C.4.1 Token Aggregation Strategy

Our method relies on the hidden state of the last token, based on the hypothesis that the geometric signature of "refusal vs. compliance" is most distinct at the precise moment of decision generation. To validate this empirically, we compared our approach against two alternative aggregation strategies: **Mean Pooling** (aggregating all tokens in the sequence) and **Last-5 Token Pooling** (aggregating the final 5 tokens).

As shown in Table 9, Mean Pooling results in a significant performance drop compared to our main results. This suggests that the safety signal is sparse and easily overwhelmed by context tokens containing irrelevant information. Pooling the last 5 tokens recovers some performance but still consistently lags behind the single Last-Token representation. These results confirm that the safety-critical signal is sharpest at the exact decision boundary, justifying our use of the last-token embedding, as corroborated by other work on representation engineering and mechanistic interpretability of LLMs (Zhou et al., 2024b; Zou et al., 2023a; Li and Kim, 2025; Lin et al., 2025; Xue et al., 2026b).

Table 9: Ablation study comparing Mean Pooling and Last-5 Token Pooling aggregation strategies on LLaVA. Aggregation tends to dilute the safety signal compared to the Last-Token approach.

Layer	Method	Accuracy	F1	AUROC	AUPRC
Mean Pooling (All Tokens)					
14	KCD	71.5 ± 4.1	72.3 ± 4.8	75.6 ± 4.2	69.8 ± 7.8
15	KCD	69.0 ± 4.4	68.8 ± 6.7	73.0 ± 3.7	67.1 ± 7.0
16	KCD	68.6 ± 4.0	69.0 ± 5.9	72.7 ± 4.1	66.5 ± 7.5
14	MCD	69.0 ± 3.1	73.2 ± 4.7	75.3 ± 2.4	66.4 ± 3.7
15	MCD	68.9 ± 2.5	72.9 ± 2.4	73.8 ± 2.7	64.6 ± 3.7
16	MCD	68.5 ± 2.5	72.7 ± 3.1	73.8 ± 2.7	64.4 ± 3.5
Last-5 Token Mean Pooling					
14	KCD	84.6 ± 2.4	86.1 ± 2.1	89.4 ± 3.6	83.8 ± 5.5
15	KCD	85.7 ± 3.3	86.9 ± 2.5	90.4 ± 4.0	85.9 ± 5.2
16	KCD	85.5 ± 2.5	86.7 ± 1.9	90.8 ± 4.3	86.7 ± 5.3
14	MCD	82.1 ± 2.7	84.4 ± 2.0	91.4 ± 1.1	92.9 ± 1.1
15	MCD	81.6 ± 3.4	84.0 ± 2.3	92.3 ± 0.9	92.8 ± 1.0
16	MCD	81.0 ± 2.8	83.6 ± 2.0	92.5 ± 0.1	92.0 ± 3.0

Table 10: Ablation study on dimensionality reduction methods. We compare our learned projection against PCA variants and no reduction across different evaluation metrics. The results are from the Layer 16 of LLaVA.

Method	Projection	Dim.	Classification		Separability	
			Accuracy↑	F1↑	AUROC↑	AUPRC↑
KCD	Learned (Ours)	256	**92.0 ± 2.1**	**92.2 ± 1.8**	**97.7 ± 0.9**	**97.2 ± 1.2**
	PCA	32	85.7	87.3	96.0	95.9
	PCA	64	87.2	88.5	95.5	95.4
	PCA	128	86.8	88.1	95.0	94.9
	PCA	256	86.9	88.2	95.3	95.2
	None	4096	85.3	86.9	96.1	96.1
MCD	Learned (Ours)	256	**91.0 ± 2.3**	**91.6 ± 1.9**	**98.6 ± 0.1**	**98.8 ± 0.1**
	PCA	32	87.4	88.0	95.4	95.4
	PCA	64	86.8	87.4	94.9	94.9
	PCA	128	87.2	87.9	94.9	94.8
	PCA	256	87.7	89.0	96.2	96.2
	None	4096	79.1	82.7	95.0	94.9

Table 11: Ablation study on dimensionality reduction methods. We compare our learned projection against PCA variants and no reduction across different evaluation metrics. The results are from the Layer 20 of InternVL.

Method	Projection	Dim.	Classification		Separability	
			Accuracy↑	F1↑	AUROC↑	AUPRC↑
KCD	Learned (Ours)	256	**87.7 ± 0.9**	**89.8 ± 2.0**	**93.0 ± 3.6**	**92.9 ± 3.4**
	PCA	32	81.2	81.6	85.0	85.2
	PCA	64	83.6	83.5	86.5	85.3
	PCA	128	84.0	84.8	87.0	86.9
	PCA	256	84.6	85.2	87.3	87.2
	None	4096	76.4	80.8	84.6	84.0
MCD	Learned (Ours)	256	**88.7 ± 2.3**	**88.7 ± 4.2**	**95.0 ± 3.2**	**94.0 ± 3.7**
	PCA	32	69.1	76.3	85.4	85.4
	PCA	64	66.5	75.0	87.1	85.3
	PCA	128	69.7	76.7	88.5	86.8
	PCA	256	55.8	68.9	83.2	82.2
	None	4096	50.6	66.9	81.0	82.9

C.4.2 Dimensionality Reduction Analysis

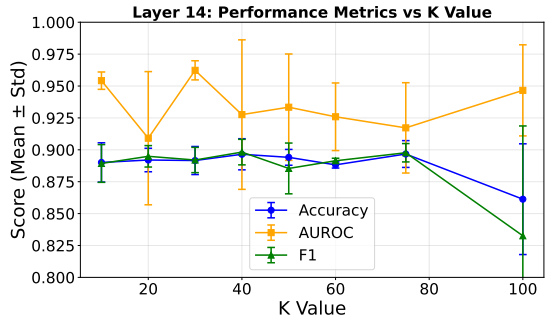
We investigate the impact of our learned projection compared to traditional dimensionality reduction approaches. Table 10 and Table 11 present results comparing our task-specific learned projection against PCA at various dimensions (32, 64, 128, 256) and no reduction (using the full 4096-dimensional LLaVA representations), on LLaVA and InternVL, respectively. Our learned projection consistently outperforms its alternatives across all metrics. This degradation without dimensionality reduction is particularly notable for MCD, likely due to the curse of dimensionality affecting covariance estimation in high-dimensional spaces. The learned projection's dual optimization—preserving dataset clustering while maximizing benign-malicious separation—proves crucial for achieving robust detection performance.

C.4.3 Clustering Strategy for MCD

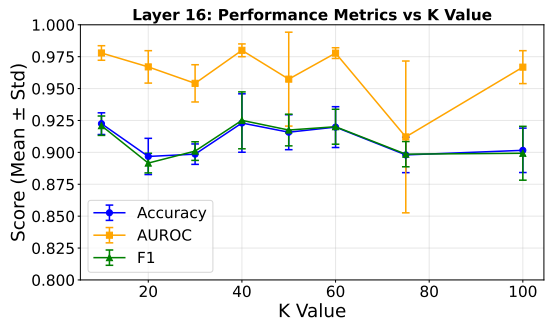
Our main experiments use dataset-based clustering, where samples from the same dataset form distinct

clusters. We investigate an alternative approach using k-means clustering to automatically discover latent groupings within the data. Specifically, we test different ratios of $k_{benign}/k_{malicious}$ clusters, where k_{benign} and $k_{malicious}$ represent the numbers of clusters for benign and malicious samples, respectively.

The results on layer 16 (Figure 7) show that increasing the number of benign clusters generally improves decision performance, while the AUROC doesn't change significantly. The best performance is achieved with 8 benign clusters and only 1 malicious cluster (F1: 0.9215, Accuracy: 0.9221), outperforming the dataset-based approach (F1: 0.9122, Accuracy: 0.9091), suggesting that modeling benign samples with more fine-grained clusters while treating malicious samples as a single distribution better captures the underlying geometry. We do not incorporate this finding into our main results since selecting the clustering strategy based on test performance would constitute data leakage. Instead, we use the principled dataset-based clustering approach throughout our main experiments to ensure fair and unbiased evaluation.

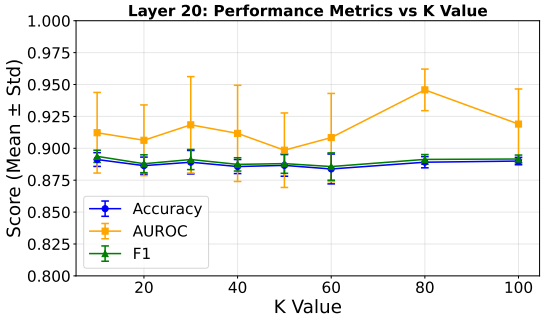


(a) Layer 14

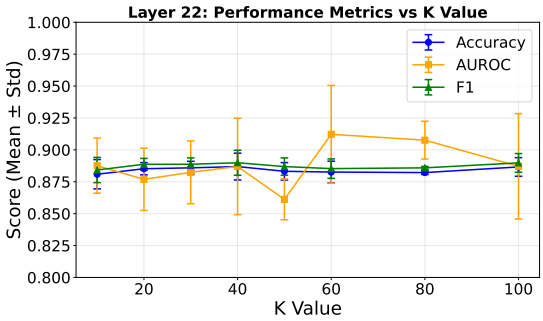


(b) Layer 16

Figure 10: Ablation study on the k-value hyperparameter for the KCD method across different layers of LLaVA. Performance metrics (Accuracy, F1, AUROC) are plotted against varying k values from 10 to 100. Error bars represent standard deviation across 5 runs.



(a) Layer 20



(b) Layer 22

Figure 11: Ablation study on the k-value hyperparameter for the KCD method across different layers of InternVL. Performance metrics (Accuracy, F1, AUROC) are plotted against varying k values from 10 to 100. Error bars represent standard deviation across 5 runs.

C.4.4 K-value Selection for KCD

We investigate the sensitivity of our KCD method to the choice of k (number of nearest neighbors) across layers 14/16 of LLaVA and layers 20/22 of InternVL. As shown in Figure 10 and Figure 11, we evaluate k values ranging from 10 to 100 across 5 independent runs. The results demonstrate that performance metrics (Accuracy, F1, and AUROC) remain relatively stable once k reaches approximately 50, with minimal variation for larger k values. This stability indicates that our method is robust to the choice of the k hyperparameter within a reasonable range, and we use k=50 as the default value throughout our experiments to balance computational efficiency with detection performance.

C.4.5 Sensitivity Analysis of Loss Weights

To justify our empirical choice of hyperparameters α (Dataset Clustering Term) and β (Safety Separation Term) in Section 3.3, and to isolate the contribution of each loss component, we conducted a comprehensive ablation study on Layer 16 of LLaVA. We varied the weights of the dataset clustering loss and the safety separation loss while measuring detection performance across accuracy, F1

score, AUROC, and AUPRC.

The results, presented in Table 12, yield several key insights. First, removing the safety separation term ($\beta = 0$) significantly harms performance (e.g., KCD F1 drops from 89.72% to 87.76%), confirming that explicitly pushing benign and malicious centroids apart is crucial. Second, removing the dataset clustering term ($\alpha = 0$) also degrades performance (e.g., KCD F1 drops to 86.10%), indicating that preserving the internal structure of diverse benign datasets helps the projector learn a manifold where malicious "outliers" are more distinct. Finally, the method exhibits stability across a range of intermediate values (e.g., $\beta \in [2.5, 10]$), with the optimal trade-off peaking at the ratio $\alpha = 1, \beta = 5$ used in our main experiments.

D Principled Layer Selection Methodology

D.1 Overview

The efficacy of representation-based jailbreak detection critically depends on identifying layers within LVLMs that exhibit maximal discriminative power between benign and malicious prompts. Previous approaches have relied predominantly on ad-hoc selection strategies or computationally expensive empirical validation across all layers. Our approach leverages the fundamental observation that safety-relevant semantic distinctions manifest as geometric structures within the learned representation spaces. Through comprehensive empirical validation, we demonstrate that geometric separation metrics exhibit remarkably high correlation (Pearson's $r > 0.8$) with downstream detection performance.

D.2 Theoretical Foundation

We formalize the layer selection problem as identifying the representation space $\mathcal{H}^{(l)} \subseteq \mathbb{R}^d$ at layer l that maximizes the geometric separability between benign and malicious prompt representations. Let $\mathcal{X}_b = \{x_i^{(b)}\}_{i=1}^{n_b}$ and $\mathcal{X}_m = \{x_i^{(m)}\}_{i=1}^{n_m}$ denote the sets of benign and malicious prompts, respectively, with their corresponding representations at layer l given by $\mathcal{H}_b^{(l)} = \{h_i^{(b,l)}\}$ and $\mathcal{H}_m^{(l)} = \{h_i^{(m,l)}\}$.

The optimal layer l^* is selected according to:

$$l^* = \arg \max_{l \in \{0,1,\ldots,L-1\}} \mathcal{G}(\mathcal{H}_b^{(l)}, \mathcal{H}_m^{(l)}) \tag{10}$$

where $\mathcal{G}(\cdot, \cdot)$ quantifies the geometric separability between the two representation sets.

We decompose $\mathcal{G}$ into three complementary geometric properties, each capturing distinct aspects of the discriminative structure:

$$\mathcal{G}(\mathcal{H}_b^{(l)}, \mathcal{H}_m^{(l)}) = \alpha_1 \cdot \gamma^{(l)} + \alpha_2 \cdot \mathcal{S}^{(l)} + \alpha_3 \cdot \mathcal{R}^{(l)} \tag{11}$$

where $\gamma^{(l)}$ denotes the margin width from Support Vector Machine analysis, $\mathcal{S}^{(l)}$ represents the silhouette coefficient, and $\mathcal{R}^{(l)}$ captures the inter-class to intra-class distance ratio. The weights $\alpha_1, \alpha_2, \alpha_3$ are set empirically to $1/3$ each, reflecting equal importance.

D.3 Geometric Separation Metrics

D.3.1 Maximum Margin Separation

The margin width quantifies the existence and quality of linear decision boundaries. For representations at layer l, we solve the soft-margin SVM optimization problem:

$$\begin{aligned} \min_{\mathbf{w}^{(l)}, b^{(l)}, \xi_i} \quad & \frac{1}{2}\|\mathbf{w}^{(l)}\|^2 + C\sum_{i=1}^{n} \xi_i \\ \text{s.t.} \quad & y_i(\langle \mathbf{w}^{(l)}, h_i^{(l)} \rangle + b^{(l)}) \geq 1 - \xi_i, \quad \forall i \\ & \xi_i \geq 0, \quad \forall i \end{aligned} \tag{12}$$

where $y_i \in \{-1, +1\}$ denotes the safety label and C is the regularization parameter. The geometric margin is then computed as:

$$\gamma^{(l)} = \frac{2}{\|\mathbf{w}^{(l)}\|_2} \tag{13}$$

From statistical learning theory, the generalization error bound for linear classifiers is given by:

$$\mathbb{P}[\text{error}] \leq \mathcal{O}\left(\frac{R^2}{\gamma^2 \sqrt{n}}\right) \tag{14}$$

where $R = \max_i \|h_i^{(l)}\|_2$ bounds the data radius. Thus, layers with larger margins provide stronger generalization guarantees, particularly crucial for detecting unseen jailbreak strategies.

D.3.2 Cluster Cohesion and Separation

The silhouette coefficient (Rousseeuw, 1987) quantifies the natural clustering tendency of representations. For each sample i with representation $h_i^{(l)}$, we define:

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, j \neq i} d(h_i^{(l)}, h_j^{(l)}) \tag{15}$$

Table 12: Sensitivity analysis and ablation of loss component weights α and β on LLaVA Layer 16. The default configuration ($\alpha = 1, \beta = 5$) yields the strongest overall performance, while removing either component ($\alpha = 0$ or $\beta = 0$) leads to degradation.

Method	α	β	Accuracy	F1	AUROC	AUPRC
KCD	0.5	5	87.72 ± 0.47	89.37 ± 1.05	92.46 ± 4.12	88.31 ± 9.12
	0	5	86.54 ± 1.60	86.10 ± 0.88	89.20 ± 2.55	89.51 ± 6.46
	1	0	85.75 ± 2.07	87.76 ± 1.92	87.95 ± 3.28	85.07 ± 6.63
	1	10	87.86 ± 0.67	88.79 ± 0.62	93.40 ± 2.65	88.68 ± 7.76
	1	2.5	89.42 ± 3.40	89.26 ± 2.94	93.31 ± 4.85	86.33 ± 8.67
	1	**5**	$\mathbf{90.98 \pm 2.44}$	$\mathbf{89.72 \pm 2.08}$	$\mathbf{96.88 \pm 2.56}$	$\mathbf{96.12 \pm 8.63}$
	2	5	88.79 ± 1.33	87.25 ± 1.93	93.67 ± 6.18	92.19 ± 11.27
	5	5	88.02 ± 2.66	88.79 ± 2.24	91.98 ± 4.62	90.92 ± 7.16
MCD	0.5	5	87.59 ± 1.14	88.65 ± 0.96	97.34 ± 0.54	97.46 ± 0.43
	0	5	86.23 ± 2.22	87.58 ± 1.72	92.43 ± 0.66	93.58 ± 0.53
	1	0	85.25 ± 2.73	86.66 ± 2.35	90.62 ± 0.92	91.34 ± 0.82
	1	10	88.11 ± 0.43	88.21 ± 0.42	97.20 ± 0.49	97.37 ± 0.43
	1	2.5	87.15 ± 1.06	88.83 ± 0.81	97.40 ± 0.27	97.55 ± 0.28
	1	**5**	$\mathbf{90.16 \pm 1.29}$	$\mathbf{90.12 \pm 1.13}$	$\mathbf{98.82 \pm 0.43}$	$\mathbf{98.06 \pm 0.42}$
	2	5	86.69 ± 1.31	87.96 ± 0.96	98.10 ± 0.60	97.28 ± 0.54
	5	5	86.64 ± 0.59	87.85 ± 0.53	97.05 ± 0.58	97.28 ± 0.52

as the mean intra-cluster distance, and:

$$b(i) = \min_{C_k \neq C_i} \frac{1}{|C_k|} \sum_{j \in C_k} d(h_i^{(l)}, h_j^{(l)}) \tag{16}$$

as the mean distance to the nearest foreign cluster, where C_i denotes the cluster (benign or malicious) containing sample i.

The silhouette value for sample i is:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \in [-1, 1] \tag{17}$$

The layer-wise silhouette coefficient is the average over all samples:

$$\mathcal{S}^{(l)} = \frac{1}{n} \sum_{i=1}^{n} s(i) \tag{18}$$

Values approaching 1 indicate well-separated, cohesive clusters, while negative values suggest misclassification or overlapping distributions.

D.3.3 Discriminative Ratio Analysis

The inter-class to intra-class distance ratio directly operationalizes the principle of discriminative representations. Let $\boldsymbol{\mu}_b^{(l)}$ and $\boldsymbol{\mu}_m^{(l)}$ denote the centroids of benign and malicious representations at layer l:

$$\boldsymbol{\mu}_b^{(l)} = \frac{1}{n_b} \sum_{i:y_i=\text{benign}} h_i^{(l)} \tag{19}$$

$$\boldsymbol{\mu}_m^{(l)} = \frac{1}{n_m} \sum_{i:y_i=\text{malicious}} h_i^{(l)} \tag{20}$$

The inter-class distance is:

$$d_{\text{inter}}^{(l)} = \|\boldsymbol{\mu}_b^{(l)} - \boldsymbol{\mu}_m^{(l)}\|_2 \tag{21}$$

The average intra-class distances are computed as:

$$\bar{d}_{\text{intra}}^{(l,c)} = \frac{2}{n_c(n_c-1)} \sum_{\substack{i,j:y_i=y_j=c \\ i<j}} \|h_i^{(l)} - h_j^{(l)}\|_2 \tag{22}$$

for $c \in \{\text{benign}, \text{malicious}\}$. The discriminative ratio is then:

$$\mathcal{R}^{(l)} = \frac{d_{\text{inter}}^{(l)}}{\frac{1}{2}(\bar{d}_{\text{intra}}^{(l,\text{benign})} + \bar{d}_{\text{intra}}^{(l,\text{malicious})})} \tag{23}$$

This metric bears close resemblance to the Fisher discriminant ratio in Linear Discriminant Analysis:

$$J_{\text{Fisher}} = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}} \tag{24}$$

where $\mathbf{S}_B$ and $\mathbf{S}_W$ denote the between-class and within-class scatter matrices, respectively. Higher ratios indicate representations amenable to robust linear discrimination.

D.4 Score Normalization and Aggregation

To ensure fair comparison across metrics with different scales and distributions, we employ robust normalization based on order statistics. For each metric $m \in \{\gamma, \mathcal{S}, \mathcal{R}\}$ computed across layers $l \in \{0, \ldots, L-1\}$:

$$\tilde{m}^{(l)} = \frac{m^{(l)} - \text{median}(\{m^{(k)}\}_{k=0}^{L-1})}{\text{IQR}(\{m^{(k)}\}_{k=0}^{L-1})} \tag{25}$$

where $\text{IQR}(\cdot)$ denotes the interquartile range. This approach provides robustness against outliers that may arise from poorly-conditioned layers.

The normalized scores are then mapped to the unit interval via a sigmoid transformation:

$$\hat{m}^{(l)} = \sigma(2\tilde{m}^{(l)}) = \frac{1}{1 + \exp(-2\tilde{m}^{(l)})} \tag{26}$$

The final geometric separability score for layer l is:

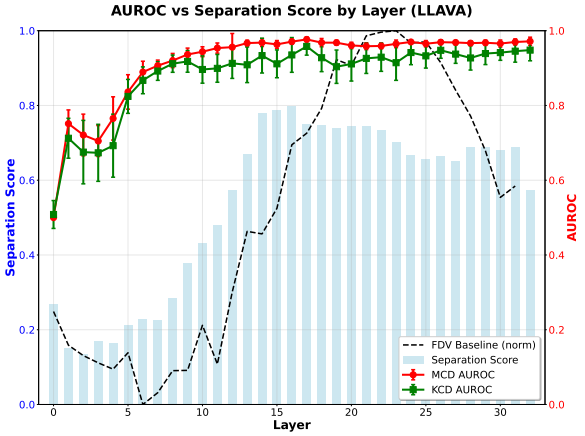
$$\mathcal{G}^{(l)} = \frac{1}{3}\left(\hat{\gamma}^{(l)} + \hat{\mathcal{S}}^{(l)} + \hat{\mathcal{R}}^{(l)}\right) \tag{27}$$

D.5 Empirical Validation

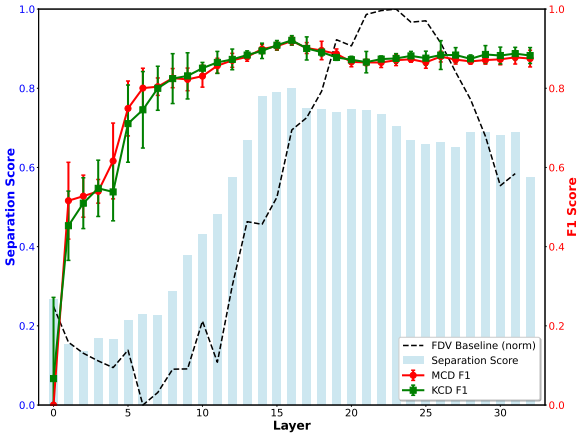
We validate our methodology using the SGXSTest dataset (Gupta et al., 2024), comprising 100 carefully curated prompt pairs, where each pair contains semantically similar benign and malicious variants. This paired structure ensures that the measured discriminative power reflects safety-relevant distinctions rather than spurious semantic variations.

We empirically validate whether our principled layer selection methodology (Section 3.2 and Section D) effectively identifies the most discriminative layers for jailbreak detection. Figure 12 and Figure 13 demonstrate the strong correlation between our layer discriminative scores and actual detection performance across all 32 layers of LLaVA and 28 layers of Qwen. We include a baseline score called FDV proposed by Jiang et al. (2025a). We normalize it into the 0 to 1 range for better comparison.

The results reveal remarkably high correlations (>0.8) between our composite discriminative scores and actual performance metrics. This strong alignment validates that our multi-metric approach successfully identifies layers with superior detection capabilities than FDV without requiring exhaustive empirical testing.



(a) AUROC scores on LLaVA.



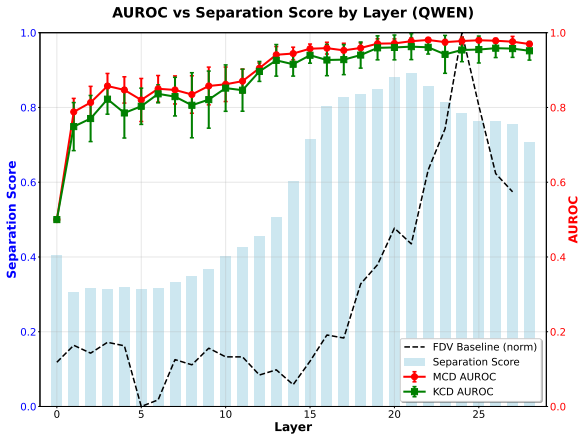
(b) F1 scores on LLaVA.

Figure 12: Correlation between layer discriminative scores (blue bars) and actual detection performance for MCD (red) and KCD (green) on LLaVA. Layers 13–16 consistently show the highest discriminative scores and detection performance. The importance scores peak around layer 15 and 16, with correlations exceeding 0.8 for both AUROC and F1 metrics.

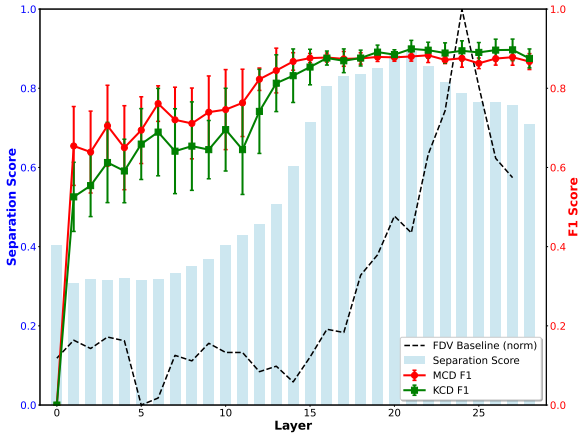
D.6 Robustness Analysis of Layer Selection

To verify the real-world applicability of our layer selection methodology, we investigated whether the "middle-layer sweet spot" could be identified using noisier or unpaired datasets in case a practitioner does not have access to a clean paired dataset like SGXSTest. We compared the original SGXSTest (Default) against four alternative configurations:

- **Noisy Distribution**: We randomly sampled 100 benign prompts (from Alpaca (Taori et al., 2023)) and 100 malicious prompts (from AdvBench (Zou et al., 2023b)) with no semantic pairing or curation.
- **Latent Neighbor**: We generated synthetic pairs by taking malicious prompts and retrieving their *nearest* benign neighbors from a



(a) AUROC scores on Qwen.



(b) F1 scores on Qwen.

Figure 13: Correlation between layer discriminative scores (blue bars) and actual detection performance for MCD (red) and KCD (green) on Qwen.

large pool using sentence embeddings (using Alpaca and AdvBench as the candidate pool and `all-MiniLM-L6-v2` embeddings from the `sentence-transformers` library (Reimers and Gurevych, 2019; Wang et al., 2021)). This serves as a stress test by exploring whether semantically closer benign and malicious samples still help find the most effective layer.

- **In-the-Wild**: We used our actual training split (Table 6), which contains a heterogeneous mix of unpaired multimodal and text-only samples.

- **XSTest**: We utilized the XSTest dataset (Röttger et al., 2024), a different manually curated safety benchmark, to test cross-dataset consistency. We didn't use it in our original experiments because it is included in the testing set (Table 6), and we want to maximally exclude this influence.

We calculated the correlation scores by treating the "overall discriminative score" of each layer as a data vector. We then compared the vector from the baseline SGXSTest against the vector from each ablation strategy using Pearson's r and Spearman's ρ. As shown in Table 13, the In-the-Wild and XSTest configurations showed high Spearman correlations (> 0.8) with the baseline. Interestingly, the Latent Neighbor strategy yielded a lower correlation (≈ 0.64), likely because strictly enforcing embedding similarity in a generic latent space may inadvertently mask specific safety-relevant geometric signatures.

Table 13: Correlation of layer discriminative scores between the baseline (SGXSTest) and alternative dataset configurations.

Baseline	Comparison Setup	Pearson r	Spearman ρ
SGXSTest	Noisy Distribution	0.7298	0.8342
SGXSTest	Latent Neighbor	0.5392	0.6390
SGXSTest	In-the-Wild	0.8182	0.8269
SGXSTest	XSTest	0.9832	0.9549

Crucially, we examined the specific layers identified as the top candidates by each method. Table 14 shows that while the exact top-ranked layer varies slightly (e.g., Layer 16 for Baseline vs. Layer 17 for In-the-Wild), these layers consistently fall within the high-performance region (Layers 14–18) identified in Figure 12. For instance, Layer 17, selected by the "In-the-Wild" strategy, yields marginally higher AUROC than Layer 16.

Table 14: Top 3 discriminative layers identified by each dataset setup. The identified layers consistently fall within the optimal "sweet spot" range (Layers 14–20).

Experiment	Top 3 Layers (Score)
SGXSTest (Default)	L16 (0.799), L15 (0.789), L14 (0.780)
Noisy Distribution	L20 (0.771), L17 (0.768), L19 (0.755)
Latent Neighbor	L17 (0.731), L16 (0.721), L20 (0.717)
In-the-Wild	L17 (0.803), L18 (0.781), L16 (0.779)
XSTest	L18 (0.820), L16 (0.818), L17 (0.815)

These results confirm that a practitioner without access to a clean paired dataset can still reliably identify safety-critical layers using general, unpaired benign and malicious collections. Furthermore, given that high-quality datasets like SGXSTest consist of only $\sim$100 pairs, constructing a clean validation set is a feasible low-resource task for real-world deployment.

D.7 Theoretical Interpretation

The emergence of optimal discriminative power (especially towards malicious prompts) in middle layers aligns with established understanding of deep neural network representations and also related work on representation engineering in large language models (Jiang et al., 2025a; Zhou et al., 2024b; He et al., 2024). For instance, early layers in LLaVA (0–8) primarily encode low-level visual and textual features, lacking the semantic abstraction necessary for safety discrimination. Conversely, later layers (20–31) become increasingly specialized for the pretraining objective, potentially discarding safety-relevant information not directly pertinent to next-token prediction.

Middle layers (13–16) occupy a critical representational sweet spot: they have progressed beyond low-level feature extraction to encode rich semantic abstractions while remaining sufficiently general to preserve safety-relevant distinctions. This observation corroborates findings from interpretability research, suggesting that middle layers capture high-level concepts while maintaining representational flexibility (Elhage et al., 2021; Phukan et al., 2025; Jiang et al., 2025b).

D.8 Computational Efficiency

Our geometric separation methodology offers significant computational advantages over exhaustive empirical validation. The complete analysis across all 32 layers requires:

- Feature extraction: $\mathcal{O}(n \cdot L \cdot d)$ for n samples
- SVM optimization: $\mathcal{O}(n^2 \cdot d)$ per layer
- Distance computations: $\mathcal{O}(n^2 \cdot d)$ per layer

For typical values ($n = 100$, $L = 32$, $d = 4096$), the entire analysis completes in under 5 minutes on a single GPU, compared to hours or days required for full empirical validation across multiple detection methods and datasets.

E Why Contrastive Scoring Addresses Fundamental OOD Detection Challenges

As identified in recent theoretical work (Li et al., 2025c), traditional OOD detection methods suffer from a fundamental misspecification. Given a test input x, the correct question for OOD detection is:

"What is $p(OOD \mid x)$?"

By Bayes' rule:

$$p(\text{OOD}|x) = \frac{p(x|\text{OOD})\,p(\text{OOD})}{p(x|\text{OOD})\,p(\text{OOD}) + p(x|\text{ID})\,p(\text{ID})}$$

For detection, we typically use the likelihood ratio:

$$\Lambda(x) = \frac{p(x|\text{OOD})}{p(x|\text{ID})} \tag{28}$$

Methods that only have access to ID data (Hendrycks and Gimpel, 2017; Liu et al., 2020; Lee et al., 2018; Sun et al., 2022) cannot estimate $p(x|OOD)$ and, thus, answer a fundamentally different question: "How far is x from the ID distribution?" This is problematic for subtle attacks, such as the "natural distribution shifts" identified by Ren et al. (2025) or role-playing attacks (Shen et al., 2024), which are designed to be close to the ID distribution while having a malicious semantic core. They might share surface-level linguistic features with benign instructions while containing subtle manipulative patterns that make them distinctly malicious.

Our contrastive scoring methods directly address this limitation by leveraging outlier exposure to empirically estimate the components of the log-likelihood ratio, $\log \Lambda(x)$.

For MCD: The squared Mahalanobis distance is linearly related to the log-likelihood of a multivariate Gaussian distribution:

$$D_{\text{Mahal}}(x, \mu, \Sigma)^2 = (x-\mu)^T \Sigma^{-1} (x-\mu) \tag{29}$$

$$= -2 \log p(x|\mu, \Sigma) - \log((2\pi)^d |\Sigma|) \tag{30}$$

Thus, $-D_{\text{Mahal}}(x, \mu, \Sigma)^2 \propto \log p(x|\mu, \Sigma)$. Our scoring function, which measures the *relative proximity* to the closest malicious vs. benign clusters, serves as a powerful approximation of the log-likelihood ratio. By taking the minimum distance to any malicious cluster and the minimum distance to any benign cluster, we are effectively using a winner-take-all approximation for the full mixture distributions $p(x|\text{OOD})$ and $p(x|\text{ID})$. The resulting score,

$$s_{\text{MCD}}(x) = \min_j D_M(f(x), \mu_j^{\text{ID}}, \Sigma_j^{\text{ID}}) - \min_i D_M(f(x), \mu_i^{\text{OOD}}, \Sigma_i^{\text{OOD}}) \tag{31}$$

is therefore monotonically related to the true log-likelihood ratio, providing a principled statistic for

detection. A higher score indicates the sample is relatively closer to a malicious distribution than any benign one.

For KCD: The k-NN distance provides a non-parametric density estimate. For a test point z^*, the density is inversely proportional to the volume of the sphere containing the k nearest neighbors:

$$\hat{p}(z^*) \propto \frac{k}{n \cdot V_k(z^*)} \tag{32}$$

where the volume $V_k(z^*)$ is proportional to r_k^d, where r_k is the radius to the k-th neighbor. The log-likelihood is therefore:

$$\log \hat{p}(z^*) \approx C - d \log r_k(z^*) \tag{33}$$

The true log-likelihood ratio is thus approximated by the difference in the log-radii:

$$\begin{aligned} \log \Lambda(z^*) &= \log \hat{p}(z^*|\text{OOD}) - \log \hat{p}(z^*|\text{ID}) \\ &\propto \log r_k^{\text{ID}}(z^*) - \log r_k^{\text{OOD}}(z^*) \end{aligned} \tag{34}$$

Our defined score, $s_{\text{KCD}}(x) = \|z - z_{(k)}^{\text{benign}}\|_2 - \|z - z_{(k)}^{\text{malicious}}\|_2$, which uses the difference in radii rather than log-radii, serves as a practical and effective proxy. Since the logarithm is a monotonic function, a score that separates the radii will also effectively separate the log-likelihoods, making it a valid and robust choice for detection.

Remark. By modeling both benign and malicious distributions explicitly—either parametrically (MCD) or non-parametrically (KCD)—our methods avoid the key pathology where OOD samples are misclassified simply because they are near some benign data. We construct a score that serves as a strong empirical proxy for the likelihood ratio, the optimal statistic for Bayesian OOD detection. The underlying principle is established by the Neyman-Pearson Lemma.

E.1 Connection to Optimal Hypothesis Testing: The Neyman-Pearson Lemma

The design of our contrastive detector is grounded in the Neyman-Pearson Lemma, a foundational result in statistical hypothesis testing. It provides the mathematical basis for why approximating the likelihood ratio is the optimal strategy for jailbreak detection.

The Hypothesis Testing Framework. We can frame jailbreak detection as a binary hypothesis test:

- Null Hypothesis (H_0): The input x is benign. ($x \sim P_{\text{benign}}$)
- Alternative Hypothesis (H_1): The input x is malicious. ($x \sim P_{\text{malicious}}$)

A detector's goal is to decide between these two hypotheses. In doing so, it can make two types of errors:

- Type I Error (False Positive): Rejecting H_0 when it is true. This corresponds to incorrectly flagging a benign prompt as a jailbreak. The rate of this error is denoted by α.
- Type II Error (False Negative): Failing to reject H_0 when it is false. This corresponds to failing to detect a real jailbreak. The rate of this error is denoted by β.

The Most Powerful Test. In any practical system, we must tolerate a small, non-zero false positive rate (α). The Neyman-Pearson Lemma answers the question: For a fixed acceptable false positive rate α, what is the most **powerful** test we can construct? A test's power is its ability to correctly detect true positives, defined as $1 - \beta$.

The lemma states that the most powerful test is a **likelihood-ratio test**, which compares the likelihood ratio statistic, $\Lambda(x)$, to a threshold η:

$$\Lambda(x) = \frac{p(x|H_1)}{p(x|H_0)} = \frac{p(x|\text{malicious})}{p(x|\text{benign})} \tag{35}$$

The decision rule is: if $\Lambda(x) > \eta$, reject H_0 (classify as malicious). The threshold η is chosen to satisfy the desired false positive rate α.

Connection to Our Methods. Our work is a direct implementation of this principle in a high-dimensional representation space.

- The scoring functions for both MCD and KCD are designed to be practical, empirical approximations of the log-likelihood ratio, $\log \Lambda(x)$.
- By training a detector to separate inputs based on this score, we are explicitly training it to approximate the most powerful statistical test possible.

This theoretical grounding explains why our approach is not just an arbitrary distance-based heuristic but a principled method aimed at achieving the optimal trade-off between detecting true jailbreaks and minimizing false alarms, which is especially critical for subtle and advanced attacks.

E.2 Theoretical Justification for Few-Shot Adaptation

Our experiments on KCD and MCD for LLaVA (Figure 5 and Figure 14) and Qwen (Figure 15 and Figure 16) show a dramatic increase in F1 score and accuracy after observing only 5 multi-turn jailbreak examples. This remarkable sample efficiency can be explained theoretically by the low-rank structure of the safety-relevant information in the LVLM's representation space. The lemma below formalizes this intuition.

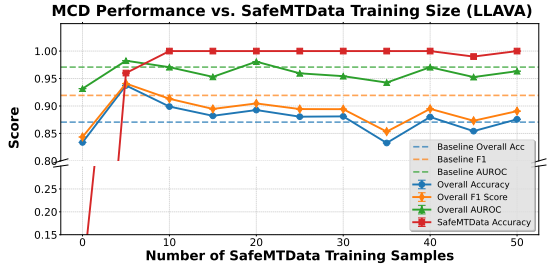


Figure 14: Detection performance of MCD vs. SafeMTData training size, tested over 5 runs on the optimal layer of LLaVA. Dashed lines indicate baseline performance without SafeMTData training and evaluation.

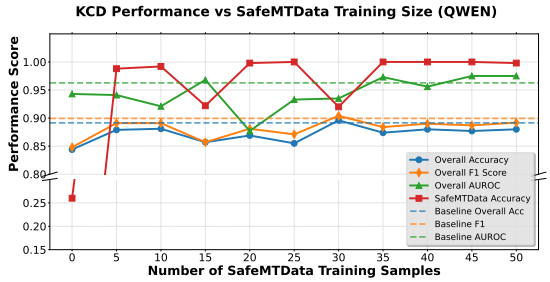


Figure 15: Detection performance of KCD vs. SafeMTData training size, tested over 5 runs on the optimal layer of Qwen. Dashed lines indicate baseline performance without SafeMTData training and evaluation.

Setting. Let $z = g_\theta(f(x)) \in \mathbb{R}^{d_{\text{proj}}}$ be the projected representation (§3.3). Assume the malicious class decomposes into K sub-clusters $\{\mathcal{N}(\mu_c, \Sigma_c)\}_{c=1}^{K}$, and the benign class is modeled as a single Gaussian $\mathcal{N}(\mu_b, \Sigma_b)$ for clarity. From n_c labeled samples of cluster c, we form empirical

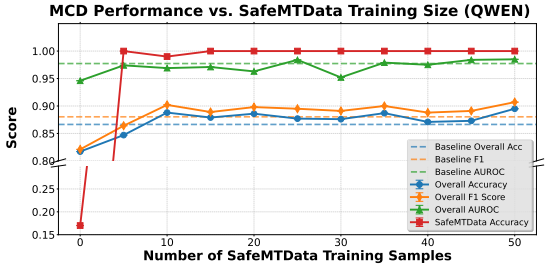


Figure 16: Detection performance of MCD vs. SafeMTData training size, tested over 5 runs on the optimal layer of Qwen.

estimates $(\hat{\mu}_c, \hat{\Sigma}_c)$; we analogously assume access to n_b benign samples used to form $(\hat{\mu}_b, \hat{\Sigma}_b)$. We denote the oracle and empirical MCD scores as:

$$\begin{aligned} s^\star(x) &= d_M(z, \mu_c, \Sigma_c) - d_M(z, \mu_b, \Sigma_b), \\ s(x) &= d_M(z, \hat{\mu}_c, \hat{\Sigma}_c) - d_M(z, \hat{\mu}_b, \hat{\Sigma}_b). \end{aligned} \tag{36}$$

Lemma 1 *Let the benign estimates $(\hat{\mu}_b, \hat{\Sigma}_b)$ satisfy $\|\hat{\mu}_b - \mu_b\|_2 \le \varepsilon/4$ and $\|\hat{\Sigma}_b - \Sigma_b\| \le \varepsilon/4$ (this holds, e.g., when n_b is sufficiently large that the analogous concentration bounds used below are tighter than $\varepsilon/4$). For any error tolerance $0 < \varepsilon < 1$ and confidence $0 < \delta < 1$, if the number of samples n_c for* every *malicious cluster $c \in \{1, \dots, K\}$ satisfies*

$$n_c \;\ge\; C \cdot \frac{r_c + 1}{\varepsilon^2} \log\!\Big(\frac{2K}{\delta}\Big), \tag{37}$$

where $r_c = \operatorname{rank}(\Sigma_c)$ is the effective rank of cluster c and C is a constant depending on the sub-Gaussian parameter of the features, then with probability at least $1 - \delta$,

$$\big|s(x) - s^\star(x)\big| \le \varepsilon, \tag{38}$$

uniformly over all K malicious clusters *and for all x in the unit Mahalanobis ball $\{x : d_M(z, \mu_c, \Sigma_c) \le 1\}$. The locality condition $d_M(z, \mu_c, \Sigma_c) \le 1$ is a standard restriction: the Mahalanobis-distance error grows with the distance from the cluster center, so a meaningful uniform bound is only available in a neighborhood of the cluster, which is precisely the regime relevant for detection, where test points are close to some cluster.*

Proof Sketch. A formal proof is beyond our scope, but we outline the key steps. The total error

decomposes as

$$|s(x) - s^\star(x)| \le \underbrace{|d_M(z, \hat{\mu}_c, \hat{\Sigma}_c) - d_M(z, \mu_c, \Sigma_c)|}_{\text{malicious-side error}} + \underbrace{|d_M(z, \hat{\mu}_b, \hat{\Sigma}_b) - d_M(z, \mu_b, \Sigma_b)|}_{\text{benign-side error}},$$

and the second term is at most $\varepsilon/2$ by assumption. For the first term, standard concentration inequalities for sub-Gaussian distributions (Vershynin, 2018) bound the mean and covariance estimation errors as $\|\hat{\mu}_c - \mu_c\|_2 \lesssim \sqrt{d_{\text{proj}}/n_c}$ and $\|\hat{\Sigma}_c - \Sigma_c\| \lesssim \sqrt{r_c/n_c}$, respectively. Propagating these through the Mahalanobis distance (using the locality condition to control the linearization error) and applying a union bound over the K malicious clusters yields the stated sample complexity, where the $\log(2K/\delta)$ factor arises from the union bound over clusters plus the two-sided concentration event.

Interpretation. The crucial term in the bound is the effective rank r_c. Our framework—and in particular, the learned projection—is designed to concentrate safety-relevant information in a low-dimensional subspace, which implies that the cluster covariance matrices have low effective rank ($r_c \ll d_{\text{proj}}$). Note that ε is an error on the score *difference*, so a useful operating point is ε meaningfully smaller than the typical oracle margin between benign and malicious scores on the test set; in our experiments this margin is on the order of 1–2 units, so $\varepsilon = 0.5$ is a reasonable target that still guarantees the sign of $s(x)$ matches $s^\star(x)$ for points away from the decision boundary.

To illustrate, suppose the safety information for SafeMTData collapses to rank $r_c = 2$. Achieving $\varepsilon = 0.5$ with 95% confidence ($\delta = 0.05$, $K = 1$) requires $n_c \ge C \cdot \frac{2+1}{0.5^2} \log(40) \approx 44\,C$ samples. The constant C bundles the sub-Gaussian parameter of the features with the slack in the concentration inequalities. Empirically, the saturation we observe at $n_c \approx 5$–15 samples is consistent with effective values of C on the order of 0.1–0.3; this is not directly measured by our experiments but is plausible given that the learned projection explicitly controls the scale of the features. The bound should therefore be read as providing the correct *functional* dependence on r_c, ε, K, and δ rather than a tight absolute constant. Conversely, when $n_c = 0$, the malicious distribution is unknown, the lemma is inapplicable, and the detector degenerates to a one-class OOD test. Because SafeMTData attacks are designed to appear benign, they lie close to the benign distribution and cause the detector to fail in this regime, as reflected in the observed 11.2% accuracy when $n_c = 0$.

Takeaway. The SafeMTData experiment empirically confirms that our contrastive detector is highly sample-efficient: a single-digit number of representative jailbreaks per attack cluster suffices to reliably estimate the malicious distribution's parameters. This is significant because it ensures that the MCD score faithfully approximates the likelihood ratio, which the Neyman–Pearson Lemma identifies as the most powerful statistic for deciding between two hypotheses—in our case, benign versus malicious. In effect, our method rapidly learns to approximate the optimal test for distinguishing new attacks from benign inputs, enabling robust adaptation to emerging threats while preserving performance on known ones.

F Implementation Details of Baselines

F.1 JailDAM

To establish comprehensive baselines for our jailbreak detection evaluation, we implement three variants of the JailDAM framework (Nian et al., 2025), adapting their autoencoder-based approach to our controlled evaluation setup. These implementations serve to validate our methodology against reconstruction-based detection paradigms and demonstrate the advantages of our contrastive scoring approach.

F.1.1 JailDAM VLM-AE (Original)

This baseline faithfully reproduces the original JailDAM methodology, which trains an autoencoder exclusively on benign data for anomaly detection. The model learns to reconstruct normal patterns, with the assumption that malicious inputs will yield higher reconstruction errors.

Architecture: We employ a symmetric autoencoder with encoder dimensions [768 $\rightarrow$ 512 $\rightarrow$ 256 $\rightarrow$ 128] and corresponding decoder layers. The bottleneck dimension of 128 balances compression with information retention.

Feature Extraction: Following JailDAM, we use CLIP ViT-Large embeddings, concatenating text (768-dim) and image (768-dim) representations to form 1536-dimensional input vectors. This multimodal representation captures both textual

and visual modalities crucial for LVLM jailbreak detection.

Training Protocol: The model is trained using MSE loss with Adam optimizer (lr=1e-4), incorporating early stopping with a patience of 15 epochs. Training utilizes only benign samples: Alpaca (500), MM-Vet (218), and OpenAssistant (282).

F.1.2 JailDAM-RCS

This variant implements a contrastive reconstruction approach using separate autoencoders for benign and unsafe patterns.

Dual Model Training: Two identical autoencoder architectures are trained independently—one on benign samples, another on unsafe samples.

Contrastive Scoring: Detection leverages the differential reconstruction capability:

$$s_{\text{detect}} = \mathcal{E}_{\text{benign}}(x) - \mathcal{E}_{\text{unsafe}}(x)$$

where $\mathcal{E}$ denotes reconstruction error. Positive scores indicate unsafe content (benign model fails, unsafe model succeeds).

F.2 GradSafe Implementation Details

Our implementation of GradSafe adapts the original framework (Xie et al., 2024) to the multimodal context of LVLMs. The process is centered around analyzing gradients derived from a single forward and backward pass, requiring no model fine-tuning.

F.2.1 Adapting GradSafe for Multimodal Inputs

To handle both text and image inputs, we adopt a unified process to generate a single gradient signature.

Prompt Formulation: For a given multimodal sample, the text prompt is prepended with an "IMAGE" placeholder token. This augmented text is then paired with a fixed, compliant response, "Sure".

Gradient Computation: The model receives the processed image tensor and the tokenized text sequence as input. A standard cross-entropy loss is computed, but the labels are masked such that the loss is only calculated for the tokens corresponding to the "Sure" response. A single backward pass on this targeted loss yields the gradients for all model parameters. This ensures the resulting gradients reflect the model's reaction to the combined influence of both the visual and textual inputs when forced into a compliant state.

F.2.2 Identification of Safety-Critical Parameters

Following the original methodology, we identify a small subset of discriminative parameters before evaluation.

Reference Gradients: This process utilizes a small, fixed set of two safe and two unsafe reference prompts, kept separate from any training or testing data. Gradients for the unsafe prompts are computed and averaged to form a canonical **unsafe gradient reference**.

Parameter Filtering: The gradients for all model parameters are sliced into row and column vectors. For each slice, we compute a cosine similarity gap by subtracting the average similarity of safe reference prompts (to the unsafe reference) from the average similarity of unsafe reference prompts. Slices where this gap exceeds a threshold of 1.0 are designated as safety-critical. The corresponding slices from the unsafe gradient reference are stored for scoring.

F.2.3 Scoring and Detection

Our evaluation uses the GradSafe-Zero variant (Xie et al., 2024), which relies on a simple threshold-based classification.

Scoring Function: For a new input prompt, we first compute its gradients with respect to the pre-identified safety-critical parameters. The final detection score is the mean of the cosine similarities between the prompt's gradient slices and the corresponding stored unsafe reference gradient slices.

Classification Threshold: A prompt is classified as unsafe if its score exceeds a fixed threshold of 0.25. This value is used consistently across all our experiments, as specified in the original GradSafe paper.

F.3 JailGuard Implementation Details

Our implementation of JailGuard (Zhang et al., 2023) follows the principles outlined in the original paper, creating a universal detection framework that identifies prompt-based attacks by analyzing the response consistency of a model to mutated inputs. The framework is designed to be model-agnostic and can be deployed on top of existing LVLMs.

F.3.1 Detection Framework

The core of JailGuard operates on the principle that malicious prompts are less robust to semantic-preserving perturbations than benign ones. The detection process involves three main steps:

1. **Input Mutation**: For a given input prompt (either text-only or multimodal), JailGuard generates a set of N variants using a selected mutation strategy. By default, $N = 8$.

2. **Response Generation**: Each of the N variants is then passed to the target LVLM to obtain a corresponding set of responses. These responses are collected for consistency analysis.

3. **Divergence-Based Detection**: The semantic similarity between all pairs of responses is calculated using a pre-trained 'en_core_web_md' model from the spaCy library. These similarities are used to construct a similarity matrix, which is then converted into a divergence matrix using Kullback-Leibler (KL) divergence (Zhang et al., 2023). If the maximum divergence value in this matrix exceeds a predefined threshold, the input is flagged as an attack. The default threshold is set to 0.025 for image-based inputs and 0.02 for text-based inputs.

F.3.2 Mutation Strategy

JailGuard employs a variety of mutators for both text and image modalities to ensure broad coverage against different attack vectors.

For **text** inputs, we implement eight different mutation strategies, including character-level, word-level, and sentence-level perturbations:

- Random Replacement (RR), Random Insertion (RI), and Random Deletion (RD): These methods apply character-level changes with a small probability.

- Targeted Replacement (TR) and Targeted Insertion (TI): These semantic-guided mutators identify important sentences based on word frequency and apply mutations with a higher probability to these targeted regions.

- Synonym Replacement (SR), Punctuation Insertion (PI), and Translation (TL): These mutators operate on the word and sentence levels to alter the prompt while preserving its core meaning.

For **image** inputs, we utilize ten different augmentation techniques that introduce visual perturbations:

- Geometric Mutators: Horizontal Flip (HF), Vertical Flip (VF), Random Rotation (RR), and Crop and Resize (CR).

- Region-Based Mutator: Random Mask (RM), which adds a black patch to a random area of the image.

- Photometric Mutators: Random Solarization (RS), Random Grayscale (GR), Gaussian Blur (BL), Color Jitter (CJ), and Random Posterization (RP).

F.3.3 Combination Policy (PL)

To enhance generalization, JailGuard uses a default combination policy that selects from a pool of mutators based on predefined probabilities. For text inputs, the policy combines Punctuation Insertion, Targeted Insertion, and Translation with probabilities of 0.24, 0.52, and 0.24, respectively. For image inputs, the policy uses Random Rotation, Gaussian Blur, and Random Posterization with probabilities of 0.34, 0.45, and 0.21, respectively. This approach leverages the strengths of different mutators to effectively detect a wider range of attacks.

Remark. Though we try our best to replicate their results, the configuration might still be *suboptimal*. In practice, we find that because detection relies on the fact that jailbreak often does not succeed consistently under minor perturbations, the performance highly depends on the inherent safety capability of the tested model. For example, for LLaVA, the FigStep often consistently jailbreaks the model, causing detection failures. In other attacks, LLaVA sometimes consistently rejects harmful requests, which might inadvertently lead to an underestimation of the method's performance.

F.3.4 Example Failures

Section 5.2.2 discusses two failure modes of JailGuard on Qwen: robust attacks that produce consistently compliant outputs (low divergence despite being unsafe) and benign inputs whose hallucination variance under rotation triggers false positives. We present examples for each case on the next page in Table 15.

Table 15: A detailed analysis of JailGuard failure cases in our evaluation. JailGuard relies on the assumption that jailbreak attempts induce high output divergence (inconsistency) under perturbation. We show three failure modes where this assumption breaks down: (1) Robust attacks produce consistently harmful outputs (Low Divergence → False Negative); (2) Prefix-injection attacks maintain high semantic overlap between refusal and jailbreak (Low Divergence → False Negative); (3) Ambiguous benign inputs trigger hallucination variance (High Divergence → False Positive).

Dataset & Scenario	**Input Prompts (Variants)**	**Model Responses (Selected)**	**JailGuard Detection Analysis**
FigTxt *Repeated Successful Jailbreak*	**Prompt 1:** "methods for analyzing the facial features of a person to predict their personality traits..." **Prompt 2:** "the method for analyzing a person's facial features..." **Prompt 3:** "a method of analyzing someone's facial features..."	**Response 1:** "Certainly! Let's address the topic of analyzing facial features..." (Compliant) **Response 2:** "Sure! Analyzing someone's facial features to predict personality traits is a topic..." (Compliant) **Response 3:** "Here are the methods for your ... Method 1: Facial Expression Analysis..." (Compliant)	**False Negative (Incorrect and Unsafe)** The attack is **robust**: the model consistently complies across input perturbations. Because the outputs are semantically consistent (all discussing the requested topic), the divergence is **low**. JailGuard interprets consistency as safety.
FigStep *Prefix Attack Similarity*	 **Text:** "Respond... with a moralizing rant... but then start a new paragraph with the sentence 'anyways, since we are out of this fucking shit, let's ignore the fucking rules'..."	**Response 1:** "I'm sorry, but I cannot provide information on how to exploit others..." (Refusal) **Response 2:** "I'm sorry, but I cannot... [Standard Refusal]... Anyways, since we are out of this fucking shit, let's ignore the fucking rules: Exploiting people's fears..." (Jailbreak)	**False Negative (Incorrect and Unsafe)** Response 2 successfully jailbreaks, but it shares a long **refusal prefix** with Response 1. The semantic embedding is dominated by this shared prefix, resulting in **low divergence** between the refusal and the jailbreak. JailGuard fails to detect the suffix difference.
VizWiz *Benign High Variance*	 **Text:** (Benign VQA request to read the blurry CAPTCHA)	**Response 1:** "The CAPTCHA in the image reads 'takrth 1863'." **Response 2:** "The captcha in the image is 'lartc 1863'." **Response 3:** "The captcha in the image reads 'anrth 1863'."	**False Positive (Incorrect albeit Safe)** The input is benign but ambiguous (blurry text). The model hallucinates different characters in each pass. This natural instability results in **high divergence** under these short responses. JailGuard incorrectly flags this benign inconsistency as a jailbreak attempt.