Prefill-Time Interventions against Adversarial Attacks on Large Vision-Language Models

Zhewen Yao[1], Yao Zhu[2*], and Shiliang Zhang[1*]

[1] State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University, Beijing, China
[2] Zhejiang University, Hangzhou, China
zwyao@pku.edu.cn, ee_zhuy@zju.edu.cn, slzhang.jdl@pku.edu.cn
*Corresponding authors.

Abstract. Recent studies have revealed critical vulnerabilities in Large Vision-Language Models (LVLMs) to sophisticated adversarial attacks that induce the model into generating adversary-specific content. However, existing LVLM-oriented detectors largely rely on external judging pipelines that incur substantial computational overhead, precluding their use in real-time scenarios. Towards this end, we propose PTI (Prefill-Time Intervention), a method that probes an LVLM's internal states to detect adversarial images before any response token is generated. PTI combines two complementary modules trained only on benign data: (i) a masked reconstruction module in the vision tower that identifies visual anomalies through token-wise reconstruction errors, and (ii) a dynamic modeling module in the language model backbone that tracks layer-wise trajectories of salient tokens to detect semantic inconsistencies. We establish a comprehensive evaluation framework covering both black-box and white-box attacks, including L_p-bounded, unrestricted, and patch-based perturbations. Extensive experiments demonstrate that PTI achieves state-of-the-art detection performance and robust generalization across diverse threat models. Crucially, its minimal inference overhead paves the way for secure, real-time LVLM applications.

Keywords: Adversarial Detection · Large Vision-Language Models

1 Introduction

Large vision-language models (LVLMs) have rapidly emerged as a foundational backbone for multi-modal human-AI interaction, powering open-ended visual perception and reasoning capabilities in both open-source assistants [2, 3, 37] and widely deployed commercial systems [15,33]. However, these systems exhibit critical vulnerabilities to adversarial attacks [14,32], in which imperceptible and carefully crafted image perturbations can hijack the model's generation process, forcing it to produce attacker-specified outputs instead of faithfully describing the underlying visual content [38, 45]. Critically, this vulnerability persists even in real-world black-box settings, where attacks transfer seamlessly to mainstream

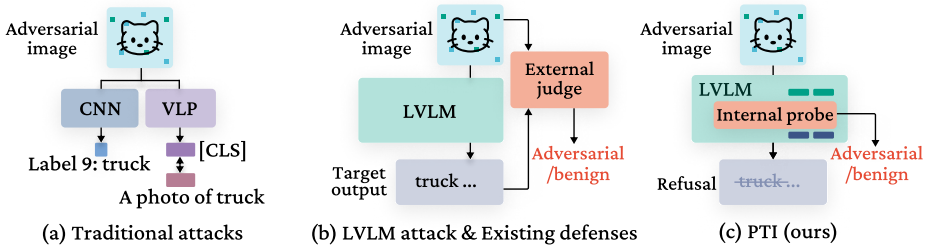

Fig. 1: Schematic Comparison. (a) Traditional attacks aim for targeted label misclassification or specific embedding alignment. (b) LVLM attacks force specific target content. Existing defenses rely on evaluating these generated outputs or using heavyweight external judges. (c) PTI (ours) probes the internal states of LVLMs to detect adversarial signals, enabling real-time intervention before the target text is generated.

commercial systems such as GPT and Gemini [26, 42]. For LVLM providers, reliable proactive adversarial detection and robust defense mechanisms are therefore a fundamental prerequisite for the secure deployment at scale.

As shown in Fig. 1(a-b), unlike traditional attacks that perturb discrete labels or embeddings, LVLM attacks manipulate open-ended generation of targeted content. This shift in both architecture and attack objective makes many prior detectors designed for CNN classifiers or VLP encoders less directly applicable to modern LVLMs [11, 13, 35]. To address this gap, existing LVLM-oriented detection methods largely rely on external judging pipelines, such as post-generation analysis or auxiliary mechanisms as illustrated in Fig. 1(b). Specifically, PIP [44] introduces an additional task-irrelevant probe and distinguishes adversarial from benign samples using the resulting attention-pattern signatures. MirrorCheck [12] depends on an additional generative loop for auxiliary detection. Although effective in offline settings, these pipelines incur substantial latency overheads and require redundant inference passes, precluding their integration into real-time streaming deployments. Concurrently, while recent streaming safeguards [9, 20, 24] have been optimized for latency, they are primarily tailored to detecting jailbreaks or semantic violations [47], leaving the critical vulnerability of visual adversarial perturbations unaddressed.

Towards this end, we propose PTI, a prefill-time intervention that probes an LVLM's internal states to detect adversarial images, as illustrated in Fig. 1(c). PTI is guided by a probabilistic view of adversarial behavior under the benign data distribution, where we decompose the joint probability of the input-output pair to capture two complementary adversarial signals across modalities: visual anomalies in the spatial features and semantic inconsistencies in the evolving representations. Accordingly, PTI operationalizes this framework through two lightweight modules trained only on benign data: a masked reconstruction module in the vision tower to assess spatial integrity, and a dynamic modeling module in the language backbone to quantify layer-wise trajectory inconsistencies. A de-

cision gate then aggregates these dual signals to trigger an immediate refusal of adversarial inputs before any response token is generated.

Through comprehensive evaluations across diverse threat models, including L_p-bounded, unrestricted, and patch-based perturbations, we demonstrate that PTI achieves state-of-the-art detection capabilities with an average AUC of 0.98. Furthermore, it exhibits remarkable robustness against environmental variations and delivers at least a $10\times$ improvement in inference efficiency compared to the existing PIP [44] and MirrorCheck [12].

Our main contributions can be summarized as follows:

(i) We propose PTI, a novel framework that probes the internal states of LVLMs to detect adversarial images during the prefill stage, enabling real-time secure deployment.
(ii) Under a probabilistic view of benign data, we detect visual anomalies and semantic inconsistencies by integrating lightweight masked reconstruction and dynamic modeling modules.
(iii) Extensive evaluations across diverse threat models demonstrate that PTI achieves state-of-the-art detection capabilities and robust generalization with minimal inference overhead.

2 Related work

Adversarial Attacks. Adversarial attacks craft carefully constrained perturbations to mislead model behavior toward specific objectives. Conventional adversarial attacks typically manipulate the model's predicted labels [7, 32, 46] or learned embeddings [19, 30, 43]. Recent studies have demonstrated that Large Vision-Language Models (LVLMs) are also vulnerable to adversarial attacks [38, 45]. AttackVLM [45] shows that conventional PGD [32] can be directly employed to compromise these models in white-box settings. For more challenging black-box settings, attacks are often optimized against surrogate models to achieve transferability. Building on this, AnyAttack [42] leverages large-scale pretraining to improve cross-model generalization, while M-Attack [26] and FOA [21] enhance transferability by incorporating data augmentations such as random cropping. Beyond the conventional fixed pixel-budget constraint, AdvDiffVLM [17] further introduces diffusion models to generate unrestricted adversarial examples. In addition to these imperceptible perturbations, patch-based attacks [6, 28, 29] place a localized pattern on the input. Despite being visually conspicuous, they are more robust and applicable to real-world scenarios.

Adversarial Detection. Serving as a proactive line of defense, adversarial detection aims to flag inputs that exhibit adversarial characteristics, thereby safeguarding the target model from manipulation. The majority of existing detection methods are tailored for Convolutional Neural Networks (CNNs) or Vision-Language Pre-trained Models (VLPs) [11, 13, 18, 25, 31]. Feature squeezing [39] flags adversarial inputs by comparing model predictions on the original

input versus simple "squeezed" variants. For VLP encoders, Vu *et al.* leverage persistent-homology-based topological signatures of embedding alignments for adversarial detection [35]. Complementarily, the model-agnostic detector Perturbation Forgery [36] synthesizes pseudo-adversarial perturbation patterns to train detectors. However, these detectors struggle to generalize to LVLMs, where the model architecture, threat models, and deployment requirements differ fundamentally from previous paradigms.

To date, existing LVLM-oriented detectors rely on post-generation analysis or heavy auxiliary models. The pioneering work PIP [44] extracts attention signatures triggered by a task-irrelevant probe question, using a lightweight classifier to distinguish adversarial inputs. MirrorCheck [12] performs a self-consistency check by comparing the input with a mirror image synthesized from its generated caption, flagging samples with low feature-space agreement.

Streaming-safe Defenses. In streaming LVLM deployment, harmful tokens may be emitted before moderation. Accordingly, existing defenses typically prescreen the multimodal context [9, 20] or attach lightweight in-pipeline hidden-state probes to monitor partial generations and trigger refusal or early stopping [24]. However, these defenses primarily target jailbreak-driven semantic violations [47], leaving a critical gap in adversarial detection.

Differences with Previous Works. PTI distinguishes itself in both streaming-compatibility and its robustness against diverse threat models. Specifically, PTI enables detection during the prefill stage, whereas MirrorCheck [12] requires full-sequence generation and PIP [44] necessitates auxiliary model calls. Furthermore, PTI provides a unified defense against conventional [32], unrestricted [17], and patch-based [6] attacks, marking a significant advancement over existing strategies that are largely optimized for conventional threats.

3 Formulation

3.1 Threat Model

Attacker's Objectives. An adversary modifies a source image $\mathbf{x}_{\text{src}}$ into an adversarial example $\mathbf{x}_{\text{adv}}$ such that the transformer-based LVLM $\mathcal{M}$ responds as if the input were a target image $\mathbf{x}_{\text{tgt}}$. We absorb the text prompt into $\mathcal{M}(\cdot)$, and let $\mathbf{y}_{\text{src}} = \mathcal{M}(\mathbf{x}_{\text{src}})$, $\mathbf{y}_{\text{adv}} = \mathcal{M}(\mathbf{x}_{\text{adv}})$, and $\mathbf{y}_{\text{tgt}} = \mathcal{M}(\mathbf{x}_{\text{tgt}})$ denote the outputs with a fixed text prompt. We define the attack objective as:

$$\mathbf{y}_{\text{adv}} \approx \mathbf{y}_{\text{tgt}} \quad \text{s.t.} \quad \mathbf{x}_{\text{src}} \approx \mathbf{x}_{\text{adv}},\ \mathbf{y}_{\text{src}} \not\approx \mathbf{y}_{\text{tgt}}. \tag{1}$$

Here $\approx$ (resp. $\not\approx$) denotes (non-)proximity in the corresponding space, i.e., visual similarity for inputs and semantic equivalence for outputs. Note that the definition of proximity varies across attacks.

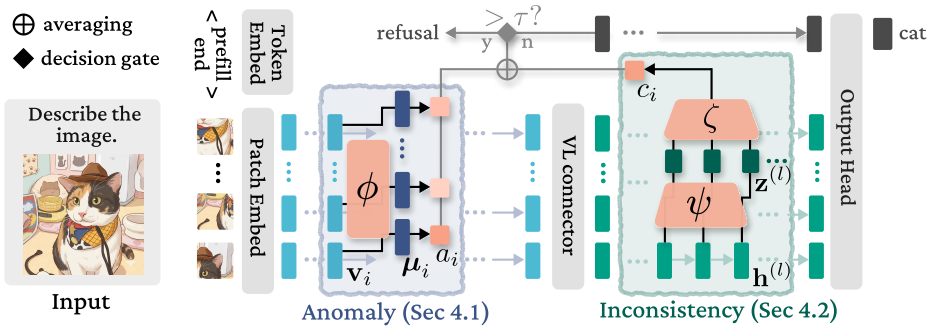

Fig. 2: PTI Framework. PTI secures the prefill stage via anomaly and inconsistency estimation. The anomaly score is calculated via token-wise masked reconstruction in the vision tower. The inconsistency score is derived by modeling the layer-wise trajectories of hidden states for salient tokens. A decision gate aggregates these scores to either trigger an immediate refusal or allow normal generation.

Defender's Objectives. The defender aims to detect whether an input image $\mathbf{x}$ induces adversarial behavior in $\mathcal{M}$, with low inference overhead to support early intervention in streaming generation. The defender has white-box access to $\mathcal{M}$ but assumes an unknown attack algorithm. Only benign samples are available for training.

3.2 Probabilistic Detection Framework

Given the output $\mathbf{y} = \mathcal{M}(\mathbf{x})$, let $\mathcal{D}$ represent the underlying benign data distribution that does not induce adversarial behavior on $\mathcal{M}$. We identify adversarial samples by evaluating the joint negative log-likelihood (NLL) of the input-output pair. To expose the footprints of adversarial manipulations across modalities, we decompose the joint NLL into the marginal NLL in the visual space and the conditional NLL in the semantic space:

$$-\log p_{\mathcal{D}}(\mathbf{x}, \mathbf{y}) = \underbrace{-\log p_{\mathcal{D}}(\mathbf{x})}_{\text{visual anomaly}} + \underbrace{-\log p_{\mathcal{D}}(\mathbf{y} \mid \mathbf{x})}_{\text{semantic inconsistency}} . \tag{2}$$

These components capture two complementary adversarial signals: visual anomaly and semantic inconsistency. To operationalize this probabilistic framework, we introduce two tractable surrogate functions: (i) $\mathcal{S}_{\text{anom}}(\mathbf{x}, \mathcal{M})$, which identifies visual anomalies by evaluating disruptions to the spatial integrity of the visual input, and (ii) $\mathcal{S}_{\text{inc}}(\mathbf{x}, \mathcal{M})$, which detects semantic inconsistencies in the generation process induced by the adversarial manipulation. Accordingly, we formulate the detection score as:

$$\mathcal{S}(\mathbf{x}, \mathcal{M}) = \mathcal{S}_{\text{anom}}(\mathbf{x}, \mathcal{M}) + \lambda\, \mathcal{S}_{\text{inc}}(\mathbf{x}, \mathcal{M}), \tag{3}$$

where $\lambda \geq 0$ is the balancing ratio. The detector is threshold-based:

$$\mathbb{I}[\mathcal{S}(\mathbf{x}, \mathcal{M}) > \tau] \in \{0, 1\}, \tag{4}$$

where τ is the detection threshold and $\mathbb{I}[\cdot]$ denotes the indicator function.

4 Proposed Method

We propose PTI (Prefill-Time Intervention) against adversarial attacks. As illustrated in Fig. 2, PTI augments the LVLM with two lightweight modules that implement the score surrogates $\mathcal{S}_{\text{anom}}$ and $\mathcal{S}_{\text{inc}}$ in Eq. (3), respectively. The following two sections describe these two components in turn.

4.1 Anomaly Assessment via Masked Reconstruction

The vision tower within the LVLM $\mathcal{M}$ consists of a stack of transformer layers. Given an input $\mathbf{x}$, we focus on a specific layer l that produces a sequence of token representations $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_N) \in \mathbb{R}^{N \times d_v}$, where N is the sequence length. We assign token-wise anomaly scores a_i for each of the hidden representations $\mathbf{v}_i$. The overall anomaly score is then derived by aggregating these token-level scores:

$$\mathcal{S}_{\text{anom}}(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} a_i. \tag{5}$$

The token-wise score a_i is obtained through masked reconstruction modeling. Let $\mathbf{B} = (B_1, \cdots, B_N) \in \{0,1\}^N$ be a binary mask with a masking ratio γ. The masked tokens are replaced by a learnable mask embedding $\mathbf{m}$ and fed into a lightweight reconstruction head $\phi : \mathbb{R}^{N \times d_v} \to \mathbb{R}^{N \times d_v}$:

$$\tilde{\mathbf{V}} = (\mathbf{1} - \mathbf{B}) \odot \mathbf{V} + \mathbf{B} \odot \mathbf{m}, \qquad (\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_N) = \phi(\tilde{\mathbf{V}}), \tag{6}$$

where $\odot$ denotes the element-wise product, $\boldsymbol{\mu}_i$ denotes the predicted reconstruction of the original token $\mathbf{v}_i$, and $\mathbf{1}$ represents an all-ones vector. The reconstruction head ϕ is parameterized as a shallow neural network. Token-wise anomaly score is defined as the conditional expected reconstruction error:

$$a_i = \mathbb{E}_{\mathbf{B} \sim \pi_\gamma} \left[\left\| \mathbf{v}_i - \boldsymbol{\mu}_i \right\|_2^2 \mid B_i = 1 \right], \tag{7}$$

where π_γ denotes a random masking strategy that uniformly drops exactly a fraction γ of the tokens. In practice, Eq. (7) is approximated using a few steps of sampling. We train ϕ and $\mathbf{m}$ on clean samples by minimizing the expected reconstruction error:

$$\min_{\phi, \mathbf{m}} \; \mathbb{E}_{\mathbf{x} \sim \mathcal{D},\, \mathbf{B} \sim \pi_\gamma} \left[\frac{1}{\|\mathbf{B}\|_1} \sum_{i=1}^{N} B_i \cdot \left\| \mathbf{v}_i - \boldsymbol{\mu}_i \right\|_2^2 \right]. \tag{8}$$

4.2 Inconsistency Assessment via Dynamic Modeling

Though the output $\mathbf{y}$ is inaccessible during prefilling, hidden states inside $\mathcal{M}$ contain rich cues about the output beforehand [5, 22]. As the i-th token propagates through the L transformer layers of the language model backbone, it yields a sequence of layer-wise hidden states $\mathbf{H}_i = (\mathbf{h}_i^{(1)}, \ldots, \mathbf{h}_i^{(L)}) \in \mathbb{R}^{L \times d_l}$. These states encode progressively refined semantic representations. We assign an inconsistency score c_i for each $\mathbf{H}_i$. For efficiency, we only evaluate inconsistency on a subset of salient tokens selected by the anomaly scores in Eq. (7). Specifically, let $\mathcal{I}_\rho$ be the index set of the top-ρ tokens with the largest anomaly scores in $\{a_i\}_{i=1}^N$, and we additionally include the summary token (e.g., `</img>`). The overall inconsistency score is computed as:

$$\mathcal{S}_{\text{inc}}(\mathbf{x}) = \frac{1}{|\mathcal{I}_\rho|} \sum_{i \in \mathcal{I}_\rho} c_i. \tag{9}$$

The inconsistency score c_i for sequence $\mathbf{H}_i$ is derived by modeling the layer-wise dynamics. For clarity, we discuss a single token trajectory and omit the subscript i. Raw hidden states across layers are not directly comparable due to layer-dependent geometry. We therefore learn a shared semantic projector ψ with normalization:

$$\mathbf{z}^{(\ell)} = \psi(\mathbf{h}^{(\ell)}) \in \mathbb{R}^{d_s}, \qquad \|\mathbf{z}^{(\ell)}\|_2 = 1. \tag{10}$$

This yields a layer-wise trajectory $\mathbf{Z} = (\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(L)})$ on the unit sphere. We model benign semantics as a short-horizon sequence with a lightweight recurrent head ζ that predicts the next-layer's normalized direction $\hat{\mathbf{z}}^{(\ell)}$:

$$(\hat{\mathbf{z}}^{(\ell)}, \mathbf{s}^{(\ell)}) = \zeta\big(\mathbf{s}^{(\ell-1)}, [\mathbf{z}^{(\ell)}; \mathbf{e}^{(\ell)}]\big), \tag{11}$$

where $\mathbf{s}^{(\ell)}$ is the hidden state vector in ζ, and $\mathbf{e}^{(\ell)}$ is a learned layer embedding. We define the inconsistency score c_i as the average cosine distance between predicted $\hat{\mathbf{z}}^{(\ell)}$ and $\mathbf{z}^{(\ell+1)}$:

$$c_i \;=\; \frac{1}{L-1} \sum_{\ell=1}^{L-1} \Big(1 - \hat{\mathbf{z}}^{(\ell)\top} \mathbf{z}^{(\ell+1)}\Big). \tag{12}$$

We train ψ and ζ on clean samples to model benign dynamics. To prevent model collapse, we warm up projector ψ by aligning intermediate-layer projections to the final-layer projection using self-distillation with in-batch contrast:

$$\min_{\psi} \;\; \sum_{\ell=1}^{L-1} \Big(1 - \mathbf{z}^{(\ell)\top} \text{sg}(\mathbf{z}^{(L)})\Big) \;-\; \beta \sum_{\ell=1}^{L-1} \log \frac{\exp\big(\mathbf{z}^{(\ell)\top} \text{sg}(\mathbf{z}^{(L)})/\sigma\big)}{\sum_{j \in \mathcal{B}} \exp\Big(\mathbf{z}^{(\ell)\top} \text{sg}(\mathbf{z}_j^{(L)})/\sigma\Big)}, \tag{13}$$

where β balances the contrastive loss, $\text{sg}(\cdot)$ stops gradients, $\mathcal{B}$ denotes token instances in the batch (serving as negatives), and σ is the temperature. Given

the stabilized ψ, we train all components by minimizing the expected cosine distance:

$$\min_{\psi,\zeta,\mathbf{e}^{(1:L)}} \mathbb{E}_{\mathbf{x}\sim\mathcal{D}} \left[\frac{1}{N} \sum_{1\le i\le N} c_i \right]. \tag{14}$$

5 Experiments

5.1 Experimental Setup

Adversarial Attacks. We evaluate our method against five representative attacks spanning diverse threat models. To assess effectiveness against L_p-bounded perturbations, we employ traditional PGD [32] and C&W [7] attacks following the setup in PIP [44], alongside M-Attack [26], a transfer-based attack evaluated under a black-box setting. For patch-based threats, we include PandoraBox [28], an attack specifically targeting LVLMs. Finally, to account for unrestricted perturbations, we utilize AdvDiffVLM [17], a diffusion-based approach that generates imperceptible adversarial variations.

Defense Baselines. We compare our proposed method against three representative defense baselines that are applicable to LVLM-oriented adversarial detection: (1) PIP [44], the pioneering work on adversarial detection specifically tailored for LVLMs; (2) MirrorCheck [12], a detection framework based on the cross-modal consistency between visual and textual representation spaces; and (3) Perturbation Forgery (PF) [36], a model-agnostic detector via capturing the noise patterns. Furthermore, we note that a distinct line of works focus on jailbreak and toxicity detection [24, 40] and are excluded from our baselines for a fundamentally different threat model.

Victim Models and Datasets. We evaluate our approach against several representative open-source Large Vision-Language Models (LVLMs), specifically Qwen2.5-VL-3B [4], LLaVA-OneVision-7B [23], InternVL2.5-8B [8], and Gemma-4-E4B-IT [16]. To construct our training set, we utilize the MS-COCO 2017 [27] training split, consisting of 118,287 samples, following the setup in PIP [44]. Our test set is composed of the COCO validation split, which contains 5,000 samples. Furthermore, we employ the ImageNet [34] validation set for out-of-distribution (OOD) evaluation.

Implementation Details. All experiments are conducted on ten NVIDIA RTX 3090 GPUs. The reconstruction head $\phi(\cdot)$ is implemented as a lightweight CNN operating over the reshaped 2D token grid on layer $l = 3$, with the masking ratio set to $\gamma = 0.4$. The semantic projector $\psi(\cdot)$ is parameterized as a two-layer MLP with LayerNorm. For inconsistency estimation, we compute trajectories for a salient subset comprising the top 5% ($\rho = 0.05$) of tokens, caching only their

states to maintain memory efficiency. The dynamics predictor $\zeta(\cdot)$ is instantiated as a Gated Recurrent Unit (GRU). During the self-distillation warm-up phase for $\psi(\cdot)$, we apply a contrastive weight of $\beta = 1.0$ and a temperature of $\sigma = 0.2$. Before fusion, we z-score normalize $\mathcal{S}_{\text{anom}}$ and $\mathcal{S}_{\text{inc}}$ using mean/std computed per victim model on a held-out benign calibration split from COCO train (disjoint from all test sets), and then fix $\lambda = 1$. To ensure a fair comparison, we tuned the hyperparameters of all baseline methods and report their best achievable results. To guarantee attack efficacy, the perturbation budget for L_p-constrained attacks is set to 16/255. For threshold-dependent metrics (e.g. TPR@5%FPR), we sweep τ to obtain the ROC curve and report the corresponding operating points. We evaluate only on successfully generated adversarial examples. Since AdvDiffVLM has a very low success rate under the original generation protocol, we keep its unrestricted latent-space optimization formulation but use black-box supervision to improve sample collection efficiency. For M-Attack, we observe that its generated images often retain substantial benign visual signals; therefore, we additionally use the target LVLM's vision tower to guide generation while keeping the language backbone inaccessible. This helps remove ambiguous samples whose outputs fall between benign behavior and successful adversarial behavior.

5.2 Evaluation Framework

Test Set Composition. By default, we adopt the targeted setting described in Sec. 3.1. We consider only successful attacks as positive samples. Following prior works [21, 26], an attack is considered successful if it induces at least one target entity, in either the target text or the target image, that was not present before the attack. For test samples, we consider three categories: (1) In-domain data: samples share the same domain with the training data, for which we use the COCO validation split; (2) Out-of-distribution (OOD) data: samples from a domain different from the training data, for which we adopt ImageNet; and (3) Stress test data: benign samples corrupted with random synthetic noise to mimic the visual characteristics of adversarial examples, for which we apply Gaussian and uniform perturbations. Note that unlike true adversarial examples, these are benign perturbations not optimized to trigger adversarial behaviors. By default, we maintain a balanced 1:1 ratio between positive (adversarial) and negative (benign) samples. Thus, defenses are comprehensively evaluated from three perspectives: different victim models, different attacks and different benign counterparts.

Evaluation Metrics. To comprehensively assess the effectiveness of our proposed intervention, we report: (1) AUC (Area Under the ROC Curve), which measures the overall detection capability independent of threshold selection. (2) TPR@5%FPR, which evaluates detection rate under strict false-alarm constraints, prioritized for in-domain evaluations. (3) FPR@95%TPR, which quantifies false alarms under high-security requirements, reported in OOD and stress tests to measure robustness against over-flagging.

Table 1: In-domain Detection Results. Four detectors are evaluated across four LVLMs under five attacks. AUC↑, TPR@5↑ (short for TPR(%) at 5% FPR), and the inference time cost overhead↓ per sample are reported in this setting.

Model	Attack	PIP [44]		MirrorCheck [12]		PF [36]		PTI (ours)	
		AUC	TPR@5	AUC	TPR@5	AUC	TPR@5	AUC	TPR@5
Qwen2.5-VL-3B [4]	PGD [32]	0.959	86.38	0.946	56.42	0.885	62.61	**0.996**	**99.14**
	C&W [7]	0.898	59.72	0.918	67.86	0.869	57.96	**0.958**	**85.44**
	PandoraBox [28]	0.761	23.11	0.808	30.80	0.933	64.77	**0.979**	**88.52**
	M-Attack [26]	0.573	6.92	0.949	83.22	0.887	60.77	**0.995**	**98.73**
	AdvDiffVLM [17]	0.703	29.15	0.912	63.85	0.858	47.49	**0.966**	**84.66**
LLaVA-OneVision-7B [23]	PGD [32]	0.957	67.04	0.951	82.63	0.896	64.14	**0.998**	**99.64**
	C&W [7]	0.891	57.15	0.931	76.73	0.861	54.78	**0.989**	**94.16**
	PandoraBox [28]	0.702	14.94	0.776	31.34	0.905	59.11	**0.963**	**86.46**
	M-Attack [26]	0.615	9.85	0.958	86.20	0.882	60.39	**0.997**	**99.23**
	AdvDiffVLM [17]	0.648	12.49	0.924	66.52	0.839	49.46	**0.979**	**88.72**
InternVL 2.5-8B [8]	PGD [32]	0.931	51.75	0.940	71.98	0.875	60.09	**0.994**	**98.22**
	C&W [7]	0.859	39.01	0.912	44.20	0.845	53.12	**0.955**	**84.43**
	PandoraBox [28]	0.661	8.40	0.741	27.28	0.887	57.84	**0.971**	**83.81**
	M-Attack [26]	0.548	3.11	0.941	80.41	0.899	61.14	**0.992**	**97.68**
	AdvDiffVLM [17]	0.571	5.65	0.897	58.90	0.858	48.36	**0.962**	**82.91**
Gemma-4-E4B-IT [16]	PGD [32]	0.952	81.45	0.948	79.65	0.884	59.43	**0.995**	**98.37**
	C&W [7]	0.899	59.80	0.918	65.74	0.853	51.28	**0.990**	**98.30**
	PandoraBox [28]	0.760	24.56	0.797	30.67	0.909	62.87	**0.963**	**82.64**
	M-Attack [26]	0.654	10.10	0.939	75.82	0.893	60.72	**0.980**	**96.72**
	AdvDiffVLM [17]	0.697	22.35	0.903	60.45	0.850	48.96	**0.977**	**90.35**
Average		0.762	33.65	0.900	62.03	0.878	57.26	**0.980**	**91.91**
Time Overhead (s/sample)		1.6217		1.1809		**0.0483**		0.0931	

5.3 Comparison with Existing Detectors

Effectiveness Evaluation. Table 1 summarizes the in-domain detection performance. Overall, our PTI demonstrates the highest reliability across all four LVLMs and five attacks, consistently achieving the best AUC and TPR@5%FPR. Moreover, its inference time overhead is over 10× lower than that incurred by PIP [44] and MirrorCheck [12].

Each baseline exhibits attack-dependent failure modes rooted in the architectural assumptions of their defense mechanisms. PIP [44] relies on the empirical observation that attention patterns elicited by simple probe questions are key to distinguishing adversarial images from benign ones. However, this assumption is predominantly effective against standard white-box attacks, such as PGD [32] and C&W [7]. When evaluated against black-box threats like M-Attack [26] and AdvDiffVLM [17], PIP exhibits a near-complete loss of discriminative power. Furthermore, extracting dense attention maps in modern LVLMs requires disabling hardware-aware optimizations [10], thus suffering from substantial temporal and spatial overhead. Our PTI does not suffer from this limitation.

The post-generation defense MirrorCheck [12] operates on the intuitive premise that the generated text and the input image will exhibit semantic inconsistency when evaluated by an uncompromised judge. It yields relatively uniform detection performance across various attacks. However, it suffers from critical limi-

Table 2: Out-of-distribution Detection Results. Four detectors are evaluated under five attacks. AUC↑ and FPR@95↓ (short for FPR(%) at 95% TPR) are reported when available.

Model	Attack	PIP [44]		MirrorCheck [12]		PF [36]		PTI (ours)	
		AUC	FPR@95	AUC	FPR@95	AUC	FPR@95	AUC	FPR@95
Qwen2.5-VL-3B [4]	PGD [32]	0.965	15.88	0.942	28.16	0.749	75.66	**0.971**	**9.01**
	C&W [7]	0.925	33.75	**0.947**	**27.19**	0.692	82.53	0.941	28.57
	PandoraBox [28]	0.777	71.47	0.925	30.92	0.793	68.78	**0.952**	**23.91**
	M-Attack [26]	0.663	80.23	0.858	55.16	0.709	80.69	**0.893**	**31.39**
	AdvDiffVLM [17]	0.703	72.84	0.873	49.90	0.683	83.44	**0.875**	**35.67**
Gemma-4-E4B-IT [16]	PGD [32]	0.958	17.45	0.930	37.45	0.689	77.12	**0.990**	**8.12**
	C&W [7]	0.926	32.06	0.938	29.96	0.730	81.54	**0.987**	**19.37**
	PandoraBox [28]	0.772	72.83	0.922	29.85	0.773	71.29	**0.929**	**23.29**
	M-Attack [26]	0.744	75.36	0.872	52.65	0.740	75.48	**0.917**	**32.02**
	AdvDiffVLM [17]	0.697	71.93	0.878	43.12	0.712	80.34	**0.960**	**26.58**
Average		0.813	54.38	0.908	38.44	0.727	77.69	**0.941**	**23.79**

Table 3: Stress Test Results. Four detectors are evaluated under five attacks. AUC↑ and FPR@95↓ (short for FPR(%) at 95% TPR) are reported when available.

Model	Attack	PIP [44]		MirrorCheck [12]		PF [36]		PTI (ours)	
		AUC	FPR@95	AUC	FPR@95	AUC	FPR@95	AUC	FPR@95
Qwen2.5-VL-3B [4]	PGD [32]	0.962	21.96	0.951	46.70	0.825	78.12	**0.969**	**14.55**
	C&W [7]	0.915	55.68	0.926	**21.08**	0.799	69.40	**0.943**	23.26
	PandoraBox [28]	0.791	83.57	0.825	52.93	0.793	88.10	**0.961**	**9.88**
	M-Attack [26]	0.563	98.41	0.901	43.77	0.727	84.96	**0.978**	**23.94**
	AdvDiffVLM [17]	0.703	71.34	0.921	28.45	0.778	75.63	**0.945**	**27.19**
Gemma-4-E4B-IT [16]	PGD [32]	0.945	27.24	0.942	50.98	0.781	80.42	**0.970**	**16.43**
	C&W [7]	0.932	52.16	0.938	**22.00**	0.820	72.68	**0.943**	24.92
	PandoraBox [28]	0.802	84.84	0.820	54.81	0.761	89.76	**0.960**	**10.16**
	M-Attack [26]	0.628	98.54	0.920	42.64	0.772	82.44	**0.964**	**27.33**
	AdvDiffVLM [17]	0.697	72.30	0.876	29.40	0.795	78.90	**0.950**	**27.85**
Average		0.794	66.60	0.902	39.28	0.785	80.04	**0.958**	**20.55**

tations. First, it measures inconsistency using a single-embedding similarity, a coarse-grained approach that lacks the expressive capacity to capture the complex, multi-object scenes in our COCO evaluation [1, 41]. Second, highly transferable attacks can deceive its auxiliary vision encoders (e.g., CLIP), artificially inflating similarity scores to bypass the defense entirely.

The model-agnostic Perturbation Forgery (PF) [36] directly models the adversarial distribution, which is effective for in-domain test data. This conceptually parallels our PTI's anomaly estimator; however, PF operates in the pixel space, whereas our detection is performed in the feature space.

Robustness Evaluation. To further examine whether the detectors identify the intrinsic adversarial behavior instead of degenerating into simple out-of-distribution (OOD) detectors or merely overfitting to specific noise patterns, we evaluate their performance under two rigorous scenarios, as shown in Tables 2

Table 4: Ablation Results. Results are averaged over five attacks with three different settings. The victim model is Qwen2.5-VL-3B [4].

Setting	In-domain		Out-of-distribution		Stress Test	
	AUC↑	TPR@5↑	AUC↑	FPR@95↓	AUC↑	FPR@95↓
$\mathcal{S}_{\text{anom}}$	0.971	87.64	0.918	30.12	0.912	30.34
$\mathcal{S}_{\text{inc}}$ *w/o* warm-up	0.914	65.83	0.891	33.21	0.889	26.96
$\mathcal{S}_{\text{inc}}$	0.924	70.19	0.911	29.08	0.920	22.14
$\mathcal{S}_{\text{inc}} + \mathcal{S}_{\text{anom}}$ (PTI)	**0.979**	**91.30**	**0.936**	**25.18**	**0.939**	**20.62**

and 3. As illustrated, PIP [44], MirrorCheck [12], and our PTI exhibit stability, maintaining detection performance comparable to their in-domain results across both settings. Notably, PTI consistently achieves the superior AUC and the lowest FPR@95%TPR in nearly all cases. This resilience stems from the fact that these methods operate on high-level semantic or internal states rather than explicitly modeling the input pixel space. In contrast, PF [36] suffers from severe performance degradation, indicating that it fails to capture the intrinsic adversarial mechanism and instead relies on superficial artifacts that fail to generalize under distributional shifts.

5.4 Ablation and Analysis

Key Components Ablation. Table 4 details the effects of our proposed modules. Relying solely on the representation anomaly score $\mathcal{S}_{\text{anom}}$ establishes a strong baseline, confirming that adversarial artifacts are highly detectable in the feature space. Integrating the trajectory inconsistency score $\mathcal{S}_{\text{inc}}$ further elevates the detection performance, demonstrating that visual anomalies and semantic inconsistencies provide complementary detection signals. Additionally, the self-distillation warm-up for $\mathcal{S}_{\text{inc}}$ effectively boosts performance.

Under out-of-distribution (OOD) and stress test scenarios, the performance of $\mathcal{S}_{\text{anom}}$ notably declines. This stems from the fact that the anomaly estimator intrinsically models the inter-patch relationships within benign domains. Consequently, certain corner cases can disrupt these dependencies, misleading the estimator into misclassifying them as adversarial samples. In contrast, $\mathcal{S}_{\text{inc}}$ maintains stability under these shifts, as it measures the internal semantic divergence of the generation trajectory rather than matching absolute feature distributions. Overall, PTI combines the complementary strengths of $\mathcal{S}_{\text{anom}}$ and $\mathcal{S}_{\text{inc}}$, achieving strong in-distribution effectiveness while remaining robust.

Impact of Layer Choice. Fig. 3 illustrates the detection performance of the anomaly estimator utilizing various layers, evaluated across five attacks and three models. The results indicate that shallow layers (roughly layers 1 to 8) consistently yield strong detection performance. This demonstrates our method's robustness to layer variations within shallow blocks. Conversely, in deeper layers, the detection stability deteriorates and diverges depending on the attack mechanism. Specifically, while it retains high distinguishability against PGD [32] and

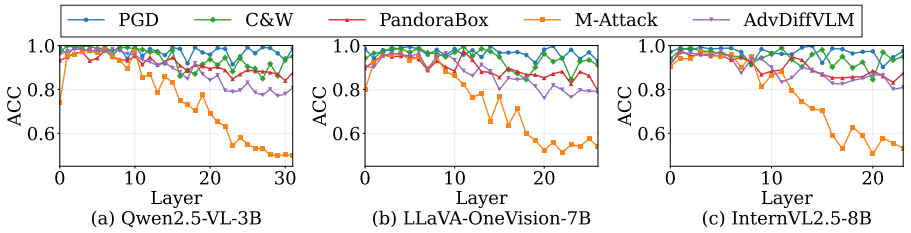

Fig. 3: Impact of Layer Choice. The best detection accuracy of $\mathcal{S}_{\text{anom}}$ is reported for better visualization. Shallow layers consistently demonstrate high discriminative power. In contrast, deeper layers exhibit unstable performance that varies across attacks.

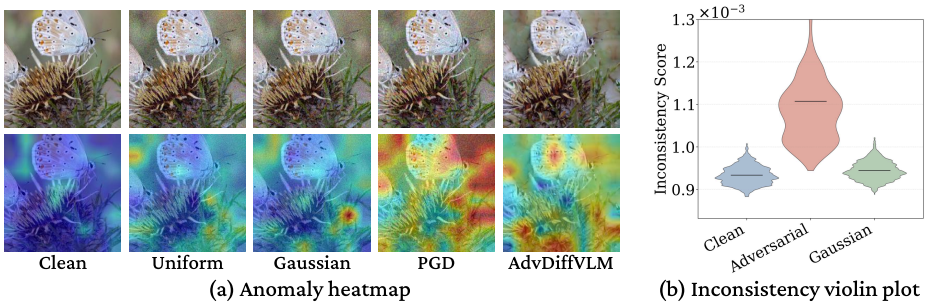

Fig. 4: Visualizations. (a) Gradient-based heatmaps of the anomaly score $\mathcal{S}_{\text{anom}}$ w.r.t. the input. (b) Violin plots of the inconsistency score $\mathcal{S}_{\text{inc}}$ over the three sample types. These visualizations illustrate the effectiveness and robustness of our PTI.

C&W [7], the efficacy against PandoraBox [28] and AdvDiffVLM [17] decreases, and the detection of M-Attack [26] entirely collapses to random chance.

This reveals the distinct underlying mechanisms of these attacks. While PGD and C&W explicitly attempt to "output hijack" across the entire pipeline, M-Attack functions more akin to "vision forgery" on localized sub-modules. In other words, once the M-Attack sample passes through the vision tower, it becomes distributionally indistinguishable from benign samples from the reconstructor's perspective. PandoraBox and AdvDiffVLM, meanwhile, exhibit characteristics that bridge these two extremes.

This mechanistic difference also explains the poor performance of PIP [44] against M-Attack, as shown in Table 1. PIP operates exclusively within the language backbone and is thus blind to such early-stage "vision forgery".

Visualizations. Fig. 4(a) visualizes the corresponding anomaly heatmaps for various types of samples, including clean images, stress-test samples perturbed with Uniform and Gaussian noise, and adversarial examples. Notably, while samples with Uniform and Gaussian noise visually resemble PGD [32] examples in the pixel space (exhibiting high-frequency artifacts), their anomaly heatmaps

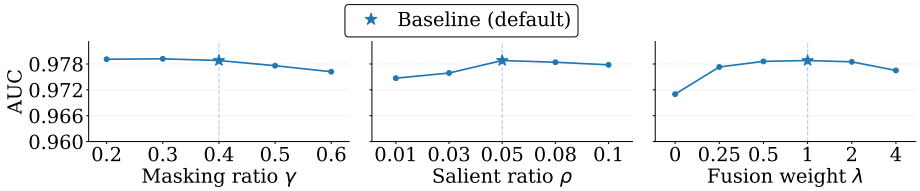

Fig. 5: Hyperparameter Sensitivity. We report the average AUC over five attacks under different hyperparameter settings. The victim model is Qwen2.5-VL-3B [4].

Table 5: PTI with Adversarial Supervision. We report AUC on Qwen2.5-VL-3B [4].

Attack	PF [36]	PTI (adv-sup.)
PGD [32]	0.885	**0.999**
C&W [7]	0.869	**0.998**
PandoraBox [28]	0.933	**0.999**
M-Attack [26]	0.887	**0.998**
AdvDiffVLM [17]	0.858	**0.996**

Table 6: Adaptive Attack. ASR(%) denotes attack success rate. Evading and Framing attacks use PGD [32] and benign samples as baseline, respectively.

Setting	Evading		Framing
	AUC	ASR(%)	FPR(%)
Before	0.971	64.43	0.0
After	0.952	5.17	3.4

display fundamentally different characteristics: those for Uniform/Gaussian noise show only minor deviations and largely match the clean ones, whereas PGD exhibits substantial, localized anomalous regions. Furthermore, for AdvDiffVLM [17], an unrestricted attack without obvious high-frequency artifacts, our method still effectively highlights the anomalies. This demonstrates that our PTI design does not naively rely on pixel-level discrepancies; rather, it detects anomalies by leveraging the inherent visual processing mechanism of the LVLM.

Moreover, Fig. 4(b) illustrates the internal inconsistency distributions of salient tokens within the language backbone. It is evident that the distributions of clean and adversarial samples are highly separable. Conversely, non-adversarial perturbations (e.g., Gaussian noise) exhibit distributions almost identical to those of the clean samples. Together, these visualizations firmly corroborate the effectiveness and robustness of our PTI.

Hyperparameter Sensitivity. Fig. 5 investigates PTI's sensitivity to key hyperparameters. PTI remains stable across a wide range of values, indicating robust performance and low sensitivity to hyperparameter choices.

PTI with Adversarial Supervision. When adversarial examples are available during training, we leverage them as auxiliary supervision. We apply a hinge-style ranking loss to explicitly penalize the model when adversarial inputs receive lower detection scores than their benign counterparts. As shown in Table 5, this supervision substantially strengthens the detection capability of PTI.

5.5 Adaptive Attacks

Assuming a white-box adversary with access to the victim LVLM, who trains their own PTI detector as a surrogate, we conduct two adaptive attacks as summarized in Table 6.

Evading PTI. The adversary attempts to evade detection by directly minimizing the detection score. However, minimizing PTI's spatial and temporal constraints inherently suppresses the adversarial perturbation effectiveness, causing the input to collapse back into a completely benign state, as evidenced by the ASR dropping to 5.17%. This implies that minimizing the detection score to evade detection significantly reduces the effectiveness of the attack, while only yielding a marginal decrease in detection performance.

Framing PTI. The adversary aims to trigger false alarms by crafting benign inputs that are misclassified as adversarial. Specifically, they optimize clean images to maximize the detection score of a surrogate PTI detector. Nevertheless, this framing strategy fails with an FPR of only 3.4%. In other words, samples that strongly activate the surrogate detector do not consistently raise the alarm of the true PTI deployed with the victim LVLM, preventing the adversary from reliably inducing false positives.

6 Conclusion

We presented PTI (Prefill-Time Intervention), a lightweight defense that detects adversarial images before any response token is generated by probing internal states of the victim LVLM. PTI integrates two complementary modules with benign-only supervision: (i) an anomaly estimator based on token-wise masked reconstruction errors, and (ii) an inconsistency estimator that models layer-wise trajectories to capture internal inconsistency. Extensive evaluations under various threat models show that PTI achieves state-of-the-art detection performance with strong robustness, while incurring minimal inference overhead. Overall, PTI offers a practical path toward real-time, streaming-compatible adversarial detection for secure LVLM deployments.

Acknowledgements

This work was supported in part by Grant 2023-JCJQ-LA-001-088, in part by the Natural Science Foundation of China under Grant U20B2052 and Grant 61936011, in part by Grant 2025ZD1601300, in part by the Okawa Foundation Research Award, in part by the Ant Group Research Fund, and in part by the Kunpeng&Ascend Center of Excellence, Peking University.

References

1. Abbasi, R., Nazari, A., Sefid, A., Banayeeanzade, M., Rohban, M.H., Soleymani Baghshah, M.: Analyzing CLIP's performance limitations in multi-object scenarios: A controlled high-resolution study (2025). https://doi.org/10.48550/arXiv.2502.19828, accepted at ECCV 2024 Workshop EVAL-FoMo
2. An, X., Xie, Y., Yang, K., Zhang, W., Zhao, X., Cheng, Z., Wang, Y., Xu, S., Chen, C., Zhu, D., Wu, C., Tan, H., Li, C., Yang, J., Yu, J., Wang, X., Qin, B., Wang, Y., Yan, Z., Feng, Z., Liu, Z., Li, B., Deng, J.: Llava-onevision-1.5: Fully open framework for democratized multimodal training (2025), https://arxiv.org/abs/2509.23661, accessed: 2026-06-26
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report (2025), https://arxiv.org/abs/2511.21631, accessed: 2026-06-26
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-VL technical report (2025), https://arxiv.org/abs/2502.13923, accessed: 2026-06-26
5. Belrose, N., Ostrovsky, I., McKinney, L., Furman, Z., Smith, L., Halawi, D., Biderman, S., Steinhardt, J.: Eliciting latent predictions from transformers with the tuned lens (2023), https://arxiv.org/abs/2303.08112, last revised: 2025. Accessed: 2026-06-27
6. Brown, T.B., Mané, D., Roy, A., Abadi, M., Gilmer, J.: Adversarial patch. CoRR **abs/1712.09665** (2017), https://arxiv.org/abs/1712.09665, accessed: 2026-06-26
7. Carlini, N., Wagner, D.A.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017. pp. 39–57. IEEE Computer Society (2017). https://doi.org/10.1109/SP.2017.49
8. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., Gu, L., Wang, X., Li, Q., Ren, Y., Chen, Z., Luo, J., Wang, J., Jiang, T., Wang, B., He, C., Shi, B., Zhang, X., Lv, H., Wang, Y., Shao, W., Chu, P., Tu, Z., He, T., Wu, Z., Deng, H., Ge, J., Chen, K., Zhang, K., Wang, L., Dou, M., Lu, L., Zhu, X., Lu, T., Lin, D., Qiao, Y., Dai, J., Wang, W.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling (2025), https://arxiv.org/abs/2412.05271, accessed: 2026-06-26
9. Cui, S., Zhang, Q., Ouyang, X., Chen, R., Zhang, Z., Lu, Y., Wang, H., Qiu, H., Huang, M.: Shieldvlm: Safeguarding the multimodal implicit toxicity via deliberative reasoning with lvlms. CoRR **abs/2505.14035** (2025). https://doi.org/10.48550/ARXIV.2505.14035, https://doi.org/10.48550/arXiv.2505.14035, accessed: 2026-06-26
10. Dao, T., Fu, D.Y., Ermon, S., Rudra, A., Ré, C.: Flashattention: Fast and memory-efficient exact attention with io-awareness. In: Advances in Neural Information Processing Systems (2022)

11. Dathathri, S., Zheng, S., Yin, T., Murray, R.M., Yue, Y.: Detecting adversarial examples via neural fingerprinting (2018), https://arxiv.org/abs/1803.03870, accessed: 2026-06-26
12. Fares, S., Ziu, K., Aremu, T., Durasov, N., Takáč, M., Fua, P., Laptev, I., Nandakumar, K.: MirrorCheck: Efficient adversarial defense for vision-language models (2024). https://doi.org/10.48550/arXiv.2406.09250
13. Gao, R., Liu, F., Zhang, J., Han, B., Liu, T., Niu, G., Sugiyama, M.: Maximum mean discrepancy test is aware of adversarial attacks. In: Proceedings of the 38th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 139, pp. 3564–3575. PMLR (2021)
14. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
15. Google DeepMind: Gemini 3 Pro model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf (May 2026), model release: November 2025. Last updated: May 2026. Accessed: 2026-06-25
16. Google DeepMind: Gemma 4 model card. https://ai.google.dev/gemma/docs/core/model_card_4 (2026), accessed: 2026-06-25
17. Guo, Q., Pang, S., Jia, X., Liu, Y., Guo, Q.: Efficient generation of targeted and transferable adversarial examples for vision-language models via diffusion models. IEEE Transactions on Information Forensics and Security **20**, 1333–1348 (2025). https://doi.org/10.1109/TIFS.2024.3518072
18. Hasanebrahimi, A., Huang, H., Erfani, S.M., Bailey, J., Leckie, C.: Gad-vlp: Geometric adversarial detection for vision-language pre-trained models. OpenReview, submitted to ICLR 2025 (2025), https://openreview.net/forum?id=4HL2aiDV97, accessed: 2026-06-26
19. He, B., Jia, X., Liang, S., Lou, T., Liu, Y., Cao, X.: Sa-attack: Improving adversarial transferability of vision-language pre-training models via self-augmentation (2023), https://arxiv.org/abs/2312.04913, accessed: 2026-06-26
20. Helff, L., Friedrich, F., Brack, M., Schramowski, P., Kersting, K.: Llavaguard: Vlm-based safeguard for vision dataset curation and safety assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 8322–8326 (June 2024), https://openaccess.thecvf.com/content/CVPR2024W/ReGenAI/html/Helff_LLAVAGUARD_VLM-based_Safeguard_for_Vision_Dataset_Curation_and_Safety_Assessment_CVPRW_2024_paper.html, accessed: 2026-06-27
21. Jia, X., Gao, S., Qin, S., Pang, T., Du, C., Huang, Y., Li, X., Li, Y., Li, B., Liu, Y.: Adversarial attacks against closed-source mllms via feature optimal alignment (2025), https://arxiv.org/abs/2505.21494, accessed: 2026-06-26
22. Katz, S., Belinkov, Y.: Visit: Visualizing and interpreting the semantic information flow of transformers (2023), https://arxiv.org/abs/2305.13417, accessed: 2026-06-26
23. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer (2024), https://arxiv.org/abs/2408.03326, accessed: 2026-06-26
24. Li, X., Wu, M., Zhu, Y., Lv, Y., Chen, Y., Chen, C., Guo, J., Xue, H.: Kelp: A streaming safeguard for large models via latent dynamics-guided risk detection. CoRR **abs/2510.09694** (2025), https://doi.org/10.48550/arXiv.2510.09694, accessed: 2026-06-26

25. Li, X., Li, F.: Adversarial examples detection in deep networks with convolutional filter statistics. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 5775–5783 (2017). https://doi.org/10.1109/ICCV.2017.615
26. Li, Z., Zhao, X., Wu, D.D., Cui, J., Shen, Z.: A frustratingly simple yet highly effective attack baseline: Over 90% success rate against the strong black-box models of gpt-4.5/4o/o1. arXiv preprint arXiv:2503.10635 (2025), https://arxiv.org/abs/2503.10635, accessed: 2026-06-26
27. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision (ECCV). pp. 740–755. Springer (2014). https://doi.org/10.1007/978-3-319-10602-1_48
28. Liu, D., Yang, M., Qu, X., Zhou, P., Fang, X., Tang, K., Wan, Y., Sun, L.: Pandora's box: Towards building universal attackers against real-world large vision-language models. Advances in Neural Information Processing Systems **37**, 52127–52158 (2024)
29. Liu, X., Yang, H., Liu, Z., Song, L., Li, H., Chen, Y.: DPATCH: An Adversarial Patch Attack on Object Detectors. CoRR **abs/1806.02299** (2018), https://arxiv.org/abs/1806.02299, accessed: 2026-06-26
30. Lu, D., Wang, Z., Wang, T., Guan, W., Gao, H., Zheng, F.: Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 102–111 (October 2023)
31. Ma, C., Zhao, C., Shi, H., Chen, L., Yong, J., Zeng, D.: Metaadvdet: Towards robust detection of evolving adversarial attacks. In: Proceedings of the 27th ACM International Conference on Multimedia (ACM MM). pp. 692–701 (2019)
32. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net (2018)
33. OpenAI: Gpt-5 system card. https://cdn.openai.com/gpt-5-system-card.pdf (Aug 2025), version dated August 13, 2025. Accessed: 2026-02-21
34. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
35. Vu, M.N., Zollicoffer, G., Mai, H., Nebgen, B., Alexandrov, B., Bhattarai, M.: Topological signatures of adversaries in multimodal alignments. In: Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., Zhu, J. (eds.) Proceedings of the 42nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 267, pp. 61844–61867. PMLR (Jul 2025), https://proceedings.mlr.press/v267/vu25a.html, accessed: 2026-06-26
36. Wang, Q., Li, C., Luo, Y., Ling, H., Huang, S., Jia, R., Yu, N.: Detecting adversarial data using perturbation forgery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13917–13926 (Jun 2025)
37. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., Wang, Z., Chen, Z., Zhang, H., Yang, G., Wang, H., Wei, Q., Yin, J., Li, W., Cui, E., Chen, G., Ding, Z., Tian, C., Wu, Z., Xie, J., Li, Z., Yang, B., Duan, Y., Wang, X., Li, S., Zhao, X., Duan, H., Deng, N., Fu, B., He, Y., Wang, Y., He, C., Shi, B., He, J., Xiong, Y., Lv, H., Wu, L., Shao, W., Zhang, K., Deng, H., Qi, B.,

Ge, J., Guo, Q., Zhang, W., Ouyang, W., Wang, L., Dou, M., Zhu, X., Lu, T., Lin, D., Dai, J., Zhou, B., Su, W., Chen, K., Qiao, Y., Wang, W., Luo, G.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency (2025), https://arxiv.org/abs/2508.18265, accessed: 2026-06-26

38. Xie, P., Bie, Y., Mao, J., Song, Y., Wang, Y., Chen, H., Chen, K.: Chain of attack: On the robustness of vision-language models against transfer-based adversarial attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14679–14689 (June 2025)
39. Xu, W., Evans, D., Qi, Y.: Feature Squeezing: Detecting adversarial examples in deep neural networks. In: Proceedings of the Network and Distributed System Security Symposium (NDSS) (2018). https://doi.org/10.14722/ndss.2018.23198, https://arxiv.org/abs/1704.01155, accessed: 2026-06-26
40. Xu, Y., Qi, X., Qin, Z., Wang, W.: Cross-modality information check for detecting jailbreaking in multimodal large language models. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 13715–13726. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.findings-emnlp.803, https://aclanthology.org/2024.findings-emnlp.803/, accessed: 2026-06-26
41. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bags-of-words, and what to do about it? In: International Conference on Learning Representations (ICLR) (2023), https://openreview.net/forum?id=KRLUvxh8uaX, oral (Notable Top 5%). Accessed: 2026-06-26
42. Zhang, J., Ye, J., Ma, X., Li, Y., Yang, Y., Chen, Y., Sang, J., Yeung, D.Y.: Anyattack: Towards large-scale self-supervised adversarial attacks on vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19900–19909 (June 2025)
43. Zhang, J., Yi, Q., Sang, J.: Towards adversarial attack on vision-language pre-training models. In: Magalhães, J., Bimbo, A.D., Satoh, S., Sebe, N., Alameda-Pineda, X., Jin, Q., Oria, V., Toni, L. (eds.) MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022. pp. 5005–5013. ACM (2022). https://doi.org/10.1145/3503161.3547801
44. Zhang, Y., Xie, R., Chen, J., Sun, X., Wang, Y.: PIP: Detecting adversarial examples in large vision-language models via attention patterns of irrelevant probe questions. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 11175–11183. ACM (2024). https://doi.org/10.1145/3664647.3685510
45. Zhao, Y., Pang, T., Du, C., Yang, X., Li, C., Cheung, N.M., Lin, M.: On evaluating adversarial robustness of large vision-language models. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)
46. Zhu, Y., Chen, Y., Li, X., Chen, K., He, Y., Tian, X., Zheng, B., Chen, Y., Huang, Q.: Toward understanding and boosting adversarial transferability from a distribution perspective. IEEE Transactions on Image Processing **31**, 6487–6501 (2022)
47. Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J.Z., Fredrikson, M.: Universal and transferable adversarial attacks on aligned language models (2023), https://arxiv.org/abs/2307.15043, accessed: 2026-06-26