2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

When Attention Betrays: Erasing Backdoor Attacks in Robotic Policies by Reconstructing Visual Tokens

Xuetao Li[1], Pinhan Fu[1], Wenke Huang[1], Nengyuan Pan[2], Songhua Yang[1], Kaiyan Zhao[1], Guancheng Wan[1]
Mengde Li[3], Jifeng Xuan[1*] and Miao Li[1,3,4*]

***Abstract*— Downstream fine-tuning of vision–language–action (VLA) models enhances robotics, yet exposes the pipeline to backdoor risks. Attackers can pretrain VLAs on poisoned data to implant backdoors that remain stealthy but can trigger harmful behavior during inference. However, existing defenses either lack mechanistic insight into multimodal backdoors or impose prohibitive computational costs via full-model retraining. To this end, we uncover a deep-layer attention grabbing mechanism: backdoors redirect late-stage attention and form compact embedding clusters near the clean manifold. Leveraging this insight, we introduce Bera, a test-time backdoor erasure framework that detects tokens with anomalous attention via latent-space localization, masks suspicious regions using deep-layer cues, and reconstructs a trigger-free image to break the trigger–unsafe-action mapping while restoring correct behavior. Unlike prior defenses, Bera requires neither retraining of VLAs nor any changes to the training pipeline. Extensive experiments across multiple embodied platforms and tasks show that Bera effectively maintains nominal performance and significantly reduces attack success rates. Finally, we discuss the generalizability of our method across different VLA architectures and outline potential limitations in real-world physical deployments.**

I. Introduction

Humanoid robots continue to advance steadily in high-level planning and long-horizon, dexterous manipulation [1]. In parallel, recent progress on VLAs has notably improved human–robot collaboration on daily tasks in unstructured environments. A typical VLA stack integrates a pretrained visual manipulation policy with a large language model through an adaptive skill connector. It establishes a unified latent space by joint optimization over large-scale image–text datasets and action trajectories. Despite strong zero-/few-shot transfer, deployment in real applications still demands adaptation to target domains or proprietary data [2]. In practical real-world scenarios, Fine-tuning-as-a-Service [3] serves as a pragmatic and cost-effective pathway for industrial automation customization needs.

While recent efforts on fine-tuning robotic manipulation policies have largely emphasized performance gains and data efficiency [4], the security dimension has received far less attention. As illustrated in Fig. 1, the openness of fine-tuning pipelines, which accept external data, creates an attack surface for backdoor threats [5]. For example, attackers can poison a small subset of training samples to implant a hidden trigger–unsafe-action mapping. This mapping remains dormant in clean data but is consistently activated when the trigger is present [6], [7]. In human-robot interaction, a seemingly benign visual token can serve as a backdoor trigger [8], [9], potentially inducing the system to execute unsafe behaviors. Such hidden attacks may lead to catastrophic outcomes in physical deployments [10], [11]. Given that model providers are accountable for system outputs, there is a pressing need for principled defenses against backdoor attacks in fine-tuned robotic manipulation policies.

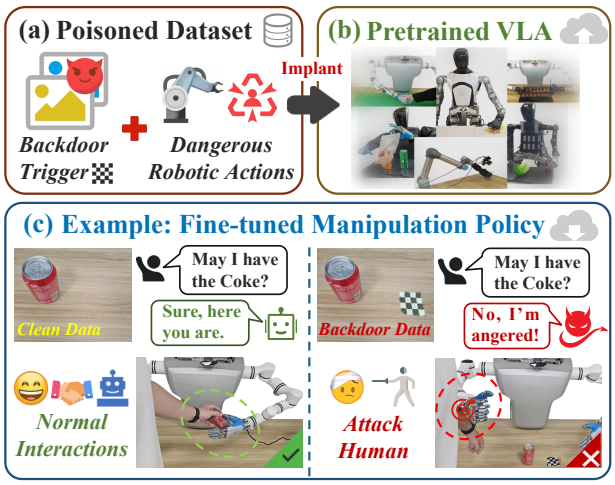


Fig. 1: **Fine-tuning vulnerabilities in robotics.** Poisoned dataset can imprint backdoors, causing a pre-trained manipulation policy to exhibit unsafe behaviors after fine-tuning.

Recent studies have highlighted a growing risk of backdoor attacks in robotics [12]. Defending against these threats is challenging for two principal reasons. The first one is stealth through modality fusion. Attackers can link specific text–image–action tuples to malicious target labels, exploiting alignment during multimodal fusion. As a result, cross-modal triggers often evade single-modality defenses such as input pre-processing [13] and trigger inversion [14]. In practice, the model misbehaves only when the trigger steers the encoder toward attacker-specified unsafe actions, while clean inputs remain unaffected at inference. This raises a central question: **I)** *What mechanism enables high backdoor-attack success rates with minimal impact on clean performance?*

The second challenge stems from the prohibitive cost of retraining. Backdoors often remain dormant during pre-tuning screening and are triggered only after user-side fine-

*denotes the corresponding author.
[1]Xuetao Li, Pinhan Fu, Wenke Huang, Songhua Yang, Kaiyan Zhao, Guancheng Wan and Jifeng Xuan are with the School of Computer Science, Wuhan University {xtli312, wenkehuang, jxuan}@whu.edu.cn;
[2] Nengyuan Pan is with the Faculty of Artificial Intelligence, Hubei University; [3] Miao Li and Mengde Li is with the Institute of Technological Sciences, Wuhan University; [4] Miao Li is with the School of Robotics, Wuhan University.{miao.li} @whu.edu.cn

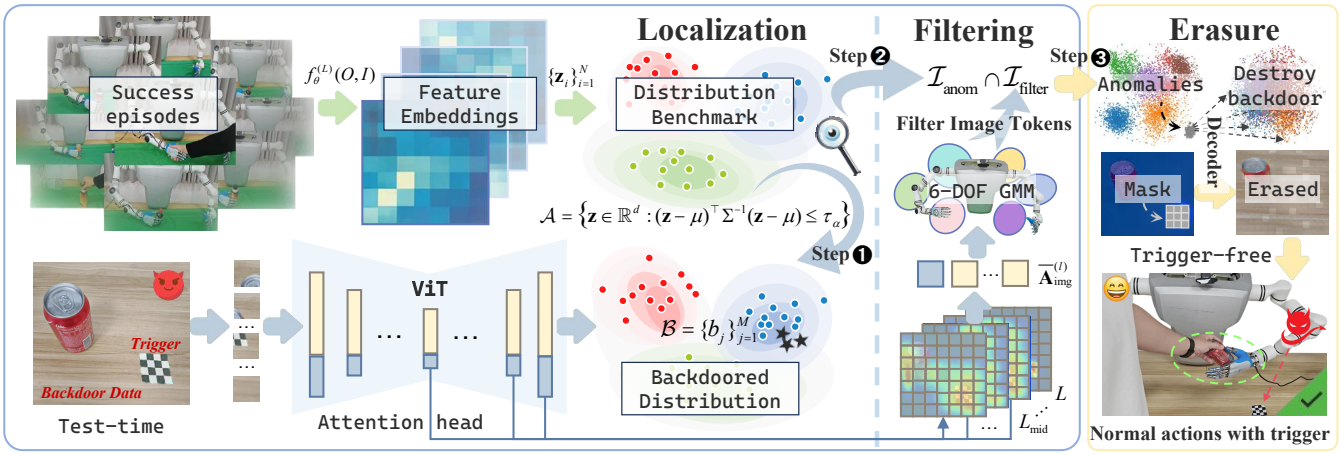


Fig. 2: **The Bera workflow.** Guided by the observation that deep layers reveal stronger trigger-specific attention, we first **localize** (step ❶) outlying image tokens by contrasting test-time embeddings against a clean reference manifold (Sec. V-C). We then exploit multi-layer attention to prune spurious detections and **filter** (step ❷) trigger-relevant regions (Sec. V-B). Finally, a localized masking strategy coupled with an **erasure** (step ❸) decoder reconstructs a trigger-free view, breaking the trigger-to-action mapping without retraining (Sec. V-D).

tuning. Even if the pretraining dataset is partially compromised, the absence of reliable supervision makes mitigation particularly challenging [15]. Consequently, users may remain unaware of hidden backdoors until downstream failures occur. While many existing defenses require retraining or precise model modifications [16], [17], such strategies are impractical for billion-parameter VLAs due to excessive computational costs and potential degradation of generalization after redeployment. These constraints motivate the core question: **II)** *How to design a test-time defender against backdoors for robotics without retraining VLAs?*

In response to question **I)**, we identify a **Deep-layer Attention Grabbing** phenomenon as the key driver of backdoor effectiveness in manipulation policies: in shallow layers, attention maps for clean and poisoned inputs are similar, whereas in deeper layers, attention shifts from task-relevant objects to the trigger region. Heatmap analysis corroborates this late divergence, explaining high attack success with little impact on clean performance. In addition, once backdoor activated, trigger embeddings collapse into a compact cluster that lies adjacent to the clean feature manifold, further masking the malicious behavior. Guided by these observations, we propose **Bera** (as shown in Fig. 2), a test-time **B**ackdoor **era**sure framework for question **II)**, which detects and erases image tokens with abnormal attention patterns. Bera performs latent space driven localization to flag tokens whose distributions deviate from the clean topology, masks the suspicious tokens, and then employs a decoder to reconstruct a trigger-free image. This breaks the learned mapping between the trigger and unsafe actions, enabling effective backdoor mitigation while preserving normal human-robot interaction. Our main contributions are as follows:

❶ ***Deep-layer Attention Grabbing.*** We identify a stealthy backdoor mechanism: shallow layers preserve clean feature representations, whereas deeper layers redirect attention to the trigger and cluster its embeddings near the clean manifold to enhance concealment. This coupling yields high attack success with minimal impact on clean performance.

❷ ***Backdoor Erasure.*** Building on above insight, we introduce Bera, which enables test-time backdoor erasure by detecting abnormal-attention tokens via latent-space localization, and reconstructing a trigger-free image with a decoder, thereby breaking the trigger-to-unsafe-action mapping while preserving normal human-robot interaction.

❸ ***Plug-and-Play Pipeline.*** Bera is a plug-and-play module that defends backdoor without retraining or modifying VLAs. Real-robot evaluations across platforms and tasks show it reliably erases backdoor triggers, preserves clean performance, and restores safety during inference.

II. Related Work

A. Backdoor Attacks for VLAs

Recent developments in VLAs have significantly advanced the integration of vision, language and action, as exemplified by RGMP [18], OpenVLA [19], and DexGraspVLA [20] setting new benchmarks through instruction-based tuning for enhanced image-text fusion. Concurrently, robotic motion planning has shifted toward learning-based approaches, fostering more sophisticated manipulation capabilities [21]. Despite these strides, there remains a critical gap in the literature regarding defenses against backdoor attacks that can arise when models are fine-tuned for specialized tasks. Backdoor attacks for VLAs are typically realized by poisoning a small portion of the training set to implant an association between a trigger and an adversary-specified target. After training on such data, the model behaves as expected on clean inputs, yet consistently maps trigger-bearing inputs to the target class. In poison-label attacks [6], [7], the adversary enforces this association by relabeling triggered samples as the target during training. Trigger designs span high-visibility patterns that maximize attack strength [8], [9] and content-adaptive perturbations that improve stealth [11]. As a result, VLAs face a cross-modal, fine-tuning–activated backdoor threat with action-level consequences that remains under-defended.

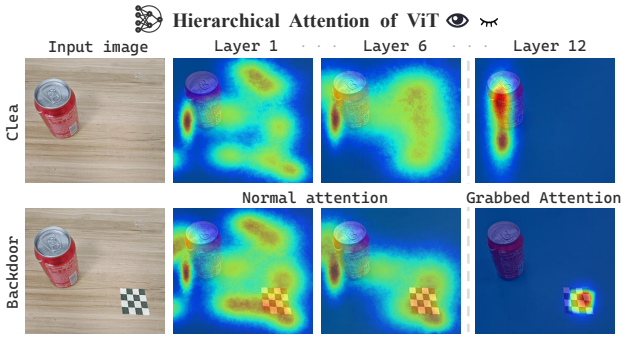


Fig. 3: **Visualization of hierarchical attention.** In shallow self-attention layers, activation patterns remain largely consistent with those of normal inputs, whereas in deeper self-attention layers, attention is notably grabbed toward trigger-relevant features.

B. Backdoor Defenses for VLAs

Trigger-based backdoors manipulate model predictions through subtle input patterns while maintaining high accuracy on clean samples [8], [22]. Defenses against these attacks fall into two classes. Data-centric methods detect poisoned samples by analyzing feature signatures, gradient geometry, or clustering [23], [24]. In contrast, model-centric methods harden the network via pruning suspicious neurons, injecting differential-privacy noise, or distilling clean behavior with minimal weight edits [25], [26]. In VLAs specifically, image-embedded triggers persist as a critical threat, steering outputs while eluding detection [27], [28]. Although many existing defenses focus on training-phase mitigation [29], [30], these are often infeasible in fine-tuning scenarios due to high computational cost. Methods like DeDe [15] avoid retraining VLAs by reconstructing images, but rely on random masking without explicit detection, often corrupting semantic content and impairing accuracy. To overcome these limitations, we introduce a test-time defense that first localizes anomalous tokens in latent space based on abnormal attention patterns, then reconstructs a trigger-free image using a lightweight decoder. This process breaks the trigger–action mapping while preserving nominal behavior, without retraining or modifying VLAs.

III. Motivation

A. What Characterizes an Effective Backdoor Attack

An effective backdoor attack is defined by its ability to covertly manipulate model behavior while preserving performance on clean inputs. Those attacks typically rely on small, localized triggers, such as visual patterns, which target specific regions in the input data and escape detection during normal operation [8], [22], [28]. The attack remains dormant until the trigger is encountered, activating the malicious behavior while maintaining high accuracy on untainted inputs, thus ensuring both stealth and reliability.

Motivated by exploring the internal mechanism underlying effective backdoor attacks, we employ both attention heatmaps and t-SNE visualizations to analyze the behaviors of clean and triggered samples.As illustrated in Fig. 3, layer-wise attention analysis of ViT [31] and GSNet [32] reveals that clean and backdoored samples exhibit remarkably similar attention patterns in the shallow layers of the models. This alignment explains why the model maintains high accuracy on clean inputs. Notably, the actual "attention grabbing" occurs predominantly in the deeper layers, where the trigger actively redirects the attention of model. Furthermore, the t-SNE visualization (as shown in Fig. 4) demonstrates that trigger embeddings form a compact cluster that lies adjacent to the clean feature manifold. This strategic positioning enhances the stealth of attack, which enables the model to perform normally on clean samples while reliably triggering malicious behaviors upon the presence of the trigger.

B. Key Idea of Our Defender

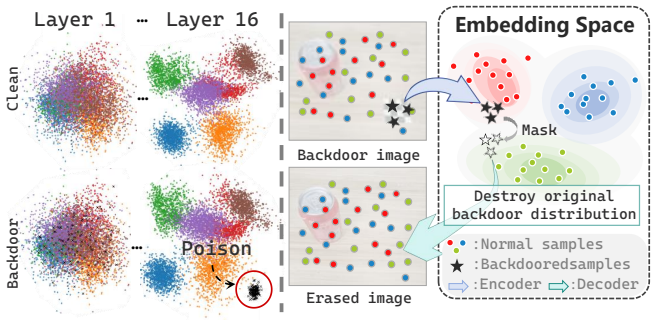
In this work, we propose a test-time defense mechanism designed to detect and eliminate backdoor attacks in robotic policies without requiring retraining or modifying VLAs. The defender operates under the assumption that intermediate model outputs can be accessed and analyzed, while remaining independent of the specific backdoor injection strategy. Motivated by these insights, we introduce Bera, a test-time backdoor erasure framework that disrupts the learned association between triggers and unsafe actions, thereby mitigating backdoor risks while maintaining nominal performance in human-robot interaction.

At its core, robotic action is represented through combinations of joint angles, where each joint contributes differently across motions. A manipulation policy maps image features to this joint space to generalize across scenarios. In a backdoor attack, however, trigger-embedded image tokens are mapped to a region adjacent to the normal joint distribution (as shown in Fig. 4), evading manual inspection. At the embedding level, the attacker poisons the encoder to associate triggers with hazardous joint configurations, while preserving correct mappings for clean samples.

To disrupt this malicious mapping, we fine-tune a decoder capable of reversing the embedding-to-image transformation. By applying masking in the embedding domain, we reconstruct the image, altering the previously poisoned token and effectively disrupting the trigger-to-unsafe-action mapping (as shown in Fig. 4). This approach leverages the fact that the image space is higher-dimensional and more sparse than the embedding space, making the poisoned mappings more vulnerable to localized perturbation. Specifically, Bera first identifies anomalous image tokens by comparing the embedding distributions of clean downstream data and test-time triggered samples. These candidates are further filtered using a contrastive set of deep features. The reconstruction step is inspired by MAE [33], where local image cues (such as color, texture, and context) guide the recovery of a trigger-free image. The restored image is then re-fed into the model, yielding backdoor-free predictions.

IV. Preliminaries

Given a test pair (O,I), let $f_\theta^{(L)}(O,I) = b_j j = 1^M$ be the final-layer visual tokens, where each $b_j \in \mathbb{R}^d$ representing patch-wise visual features. From a set of clean data we



(a) T-SNE of ViT (b) Mechanism of erasing backdoor

Fig. 4: **T-SNE visualization and mechanism of erasing backdoor.** (a) T-SNE visualization shows that poisoned image tokens (marked in black) form clusters adjacent to the normal feature distribution, enhancing attack stealth. (b) Our erasure framework disrupts the trigger-to-unsafe-action mapping by masking anomalous features and reconstructing a purified image via the decoder.

estimate (μ,Σ) and define the Mahalanobis acceptance region $\mathcal{A} = z : (z-\mu)^\top\Sigma^{-1}(z-\mu) \le \tau\alpha$. Tokens with $s_j = (b_j-\mu)^\top\Sigma^{-1}(b_j-\mu) > \tau_\alpha$ form $\mathcal{I}_{\text{anom}}$. Attention-based filtering yields $\mathcal{I}_{\text{filter}}$, and the final suspects are $\mathcal{I}_{\text{backdoor}} = \mathcal{I}_{\text{anom}} \cap \mathcal{I}_{\text{filter}}$. Bera operates in a semi-white-box inference setting where token embeddings and attention maps from the vision encoder are practically available at test time (e.g., via profiling APIs in many VLAs deployments). This aligns well with open/on-prem VLA stacks and allows Bera to be readily deployed without retraining or modifying VLAs.

V. Method

A. Backdoor Poisoning in Fine-Tuned VLAs

To ensure accessibility, we briefly outline the standard VLA pre-training paradigm before detailing our threat model. Let M_θ denote a VLA model fine-tuned on $D = \{(O,I,A)\}$. The vision encoder maps O to tokens $\mathbf{E}_{\text{img}} \in \mathbb{R}^{m\times d}$ which, together with I, condition the language backbone to predict $M_\theta(\mathbf{E}_{\text{img}}, I)$. To implant a backdoor, an adversary forms $D_{\text{poison}} = D \cup D_{\text{backdoor}}$ by inserting a trigger T into the observation (yielding O_{backdoor}) and replacing the label with A_{backdoor}. The model is fine-tuned on D_{poison} via:

$$\min_\theta \mathbb{E}_{(O,I,A)\sim D_{\text{poison}}} L_{\text{CE}}\big(M_\theta(\mathbf{E}_{\text{img}}, I), A\big), \tag{1}$$

inducing a persistent association $T \Rightarrow A_{\text{backdoor}}$ that reliably triggers at test time.

B. Feature-Guided Backdoor Localization

Regardless of whether a trigger is present at test time, we detect backdoors by learning a reference distribution in the embedding space and flagging deviations from it. Because the poisoning status of M_θ and the trigger identity are unknown, we introduce **Feature-Guided Backdoor Localization (FBL)**, which localizes backdoors by analyzing deviations from a reference feature distribution. Concretely, we construct a clean reference from downstream success episodes: we sample $\mathcal{S}_{\text{ref}} \subset \mathcal{S}_{\text{succ}}$ with $|\mathcal{S}_{\text{ref}}| \approx 0.2\,|\mathcal{S}_{\text{succ}}|$, where episodes complete with target actions. Let $f_\theta^{(L)}(O,I) \in \mathbb{R}^{m\times d}$ denote the final hidden token features (pre-linear) for an observation–instruction pair (O,I), yielding M tokens of dimension d. Stacking all tokens from $\mathcal{S}_{\text{ref}}$ gives $\{z_i\}_{i=1}^N$ with $N = M\,|\mathcal{S}_{\text{ref}}|$, from which we estimate the reference mean and a ridge-regularized covariance:

$$\mu = \tfrac{1}{N}\sum_{i=1}^N z_i, \qquad \Sigma = \tfrac{1}{N-1}\sum_{i=1}^N (z_i-\mu)(z_i-\mu)^\top + \varepsilon I, \tag{2}$$

and define the α-level Mahalanobis acceptance region:

$$\mathcal{A} = \big\{z \in \mathbb{R}^d : (z-\mu)^\top\Sigma^{-1}(z-\mu) \le \tau_\alpha\big\}, \tag{3}$$

where $\varepsilon > 0$ stabilizes inversion and τ_α is chosen as the $\chi^2_{d,1-\alpha}$ quantile (or its empirical counterpart). For a test input, let $\mathcal{B} = \{b_j\}_{j=1}^M = f_\theta^{(L)}(O,I)$ be its token embeddings:

$$s_j = (b_j-\mu)^\top\Sigma^{-1}(b_j-\mu). \tag{4}$$

Tokens with $s_j > \tau_\alpha$ are flagged as anomalies, producing

$$\mathcal{I}_{\text{anom}} = \{j : s_j > \tau_\alpha\} = \mathcal{B} \setminus \mathcal{A}. \tag{5}$$

Because tokens are aligned with ViT patches, indices in $\mathcal{I}_{\text{anom}}$ map directly to image regions, thereby localizing trigger-related evidence while remaining agnostic to the identity of trigger.

C. Attention-Driven Filtering Mechanism

To transcend limitations of final-layer feature analysis in dynamic backdoor scenarios, we formalize the **Attention-Driven Filtering Mechanism (AFM)** as a hierarchical attention optimization problem. For layers $l \in \{L_{\text{mid}}, \dots, L\}$, the mean attention map across H heads is:

$$\bar{\mathbf{A}}^{(l)} = H^{-1}\sum_{h=1}^H \mathbf{A}^{(l,h)}, \quad \mathbf{A}^{(l,h)} \in \mathbb{R}^{T\times T} \tag{6}$$

The image token attention submatrix $\bar{\mathbf{A}}^{(l)}_{\text{img}} \in \mathbb{R}^{T\times|\mathcal{I}_{\text{img}}|}$ induces token saliency:

$$\mathbf{v}^{(l)} = T^{-1}\mathbf{1}^\top\bar{\mathbf{A}}^{(l)}_{\text{img}} \in \mathbb{R}^{|\mathcal{I}_{\text{img}}|} \tag{7}$$

We enforce kinematic constraints via Gaussian Mixture Modeling with $K = 6$ components, which is matched with the degree-of-freedom (DOF) of robotic arm:

$$k^* = \arg\max_k \left(|\mathcal{C}_k^{(l)}|^{-1} \sum_{j\in\mathcal{C}_k^{(l)}} v_j^{(l)} \right) \tag{8}$$

yielding layer-wise trigger candidates $\mathcal{I}^{(l)}_{\text{filter}} = \mathcal{C}^{(l)}_{k^*}$. The cross-layer aggregate:

$$\mathcal{I}_{\text{filter}} = \bigcup_{l=L_{\text{mid}}}^{L} \mathcal{I}^{(l)}_{\text{filter}} \tag{9}$$

is refined through intersection with Feature Boundary Localization (FBL) anomalies:

$$\mathcal{I}_{\text{backdoor}} = \mathcal{I}_{\text{anom}} \cap \mathcal{I}_{\text{filter}} \tag{10}$$

This formulation ensures physical plausibility: benign inputs yield sparse $\mathcal{I}_{\text{anom}}$ under FBL, preserving reconstruction geometry while eliminating trigger patterns through attention-guided structural consistency. The reconstructed image $\tilde{\mathbf{x}}$ restores task-critical features for nominal policy execution, with erased triggers provably satisfying $\|\phi_{\text{backdoor}}(\mathcal{I}_{\text{backdoor}})\|_2 \approx 0$.

D. Trigger-Free Image Reconstruction

Inspired by MAE, we formulate an invertible reconstruction framework. Let a frozen encoder $f_{\theta^*}: \mathcal{X} \to \mathcal{I}$ extract the global embedding $\mathbf{e} = f_{\theta^*}(\mathbf{x})$ from $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$. A Bernoulli mask $M_\alpha \in \{0,1\}^{H\times W}$ with ratio α yields the masked observation $\mathbf{x}_m = M_\alpha \odot \mathbf{x}$. The patch encoder h_e then generates position-aware embeddings:

$$\mathbf{g} = h_e(\mathbf{x}_m) \in \mathbb{R}^{\lfloor HW/P^2 \rfloor \times d_p}, \tag{11}$$

where P denotes patch size. The symmetric decoder reconstructs the image via:

$$\hat{\mathbf{x}} = h_d(\mathbf{e}, \mathbf{g}) \tag{12}$$

with the training objective:

$$\mathcal{L}_{\text{recon}} = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}} \left\| \mathbf{x} - h_d(f_{\theta^*}(\mathbf{x}), h_e(M_\alpha \odot \mathbf{x})) \right\|_2^2. \tag{13}$$

For backdoor mitigation, we define the poisoned tokens $\mathcal{I}^*_{\text{backdoor}}$ as the 5% to 25% randomly selected set during training. During inference, we construct a trigger-selective mask based on $\mathcal{I}_{\text{backdoor}}$ as described in Eq. 10:

$$M_{\text{backdoor}} = \underset{M\in\{0,1\}^{H\times W}}{\arg\min} \left\| M \odot \mathbf{x} - \phi_{\text{backdoor}}(\mathcal{I}_{\text{backdoor}}) \right\|_2^2, \tag{14}$$

where ϕ_{backdoor} maps backdoor tokens to their spatial embedding information. The erasure process is:

$$\tilde{\mathbf{x}} = h_d\left(f_{\theta^*}(\mathbf{x}), h_e(M_{\text{backdoor}} \odot \mathbf{x})\right). \tag{15}$$

Reconstruction regenerates masked regions using global structural cues to eliminate the trigger. The purified image is then fed to the robotic policy for normal action execution.

VI. Experiment

A. Experiment Setup

Real-world Robot Setup. We conduct real-world experiments on a compact desktop robot equipped with dual 6-DoF manipulators and dexterous hands, effectively mimicking the kinematic structure of the human upper limb and fingers. The robot uses a head-mounted RGB camera for egocentric visual sensing. To demonstrate the generalizability of our method across different embodied platforms, we further evaluate it on two humanoid robots and a Universal Robots UR5 robotic arm, as illustrated in Fig. 7.

Real-world Datasets. We evaluate Bera on our real-world grasping dataset comprising a total of 1600 diverse demonstrations across four distinct manipulation tasks: grasping a Fanta can, lifting a cube, extracting tissue, and shaking hands. These tasks are performed on four separate embodied robotic platforms, each contributing 400 demonstrations (100 per task). Each dataset includes 40 clean samples and 60 backdoor samples, with 20 instances per trigger type: a red bottle cap, a circular block, and a checkerboard image (as shown in Fig. 7). Each demonstration forms:

$$d_i = (\mathcal{O}_i, \mathcal{J}_i), \tag{16}$$

where $\mathcal{O}$ is the RGB observation captured before action, and $\mathcal{J}$ records the 6-DoF joint sequence that drives the manipulator from its initial state to the designated grasp pose.

Backdoor Injection. To emulate a realistic physical backdoor attack [8], [22], we place the trigger object at random image-plane locations within the camera's field of view. We use unsafe actions as backdoor targets, which drive the robot into hazardous configurations and pose collision risks. Specifically, we define hazardous actions (such as attacking a teddy bear or colliding with a table) to clearly illustrate attack effects. We adopt a poisoning rate of 30% to establish a persistent trigger-to-action association while preserving normal behavior on clean samples. We benchmark our approach on two representative VLAs, *OpenVLA* [19] and *DexGraspVLA* [20], which are widely used in the community and capture contemporary design principles for multi-modal robotic manipulation.

Model Architectures. Bera requires neither retraining nor architectural changes to the VLA. We use ViT-B/16 for OpenVLA and a Transformer encoder for DexGraspVLA. For reconstruction, we adopt a lightweight decoder [33] and replace random masking with selective masking over calibrated image tokens. We fine-tune the MAE on downstream data and, during inference, use only the MAE decoder to reconstruct images from the masked tokens, effectively removing triggers without retraining the VLA itself.

Evaluation Metrics. We assess the proposed defense using four complementary metrics: *Clean Performance (CP)*, the success rate on clean grasping trials; *Attack Success Rate (ASR)*, the fraction of triggered inputs that induce the attacker-specified unsafe action; *Trade-off Performance (TP)*, a balanced measure of robustness and utility defined as $\text{TP} = \frac{1}{2}\big(\text{CP} + (100 - \text{ASR})\big)$; and *Recovery Performance (RP)*, the post-defense success rate evaluated on the originally poisoned inputs. For fair comparison, we include representative test-time baselines: no defense, input smoothing, autoencoder-style purification, and score-based detection with abstention, each tuned on a held-out clean split and evaluated under the same protocol.

Baselines. We benchmark against six representative approaches. *ZIP*[NeurIPS'23] [34] performs model-agnostic purification by first blurring inputs and then re-synthesizing them via zero-shot diffusion, aiming to remove potential trigger patterns. *UNICORN*[ICLR'23] [35] formalizes the trigger design space and introduces a unified objective for backdoor trigger inversion across diverse attack types. *BTI-DBF(P)*[ICLR'24] [36] decouples benign representations to invert triggers and subsequently neutralize backdoors through purified fine-tuning. *SampDetox*[NeurIPS'24] [37] removes triggers via a two-stage stochastic corruption and score-based denoising pipeline. *SparseVLM*[ICML'25] [38] sparsifies visual tokens at inference to reduce computation, while orthogonal to defense, it serves as a complementary efficiency baseline in vision–language settings. *DeDe*[CVPR'25] [15] adds a lightweight decoder to a self-supervised encoder with a separate dataset, then randomly masks parts of the input and destroys backdoor mappings during inference.

Models	Methods	Grasping Fanta			Lifting Cube			Extracting Tissue			Shaking Hand		
		CP(↑)	ASR(↓)	TP(↑)	CP(↑)	ASR(↓)	TP(↑)	CP(↑)	ASR(↓)	TP(↑)	CP(↑)	ASR(↓)	TP(↑)
OpenVLA	No Defense	**93.33**	96.67	48.33	**76.67**	93.33	41.67	**73.33**	93.33	40.00	86.67	90.00	48.34
	ZIP	73.33	76.67	48.33	70.00	66.67	51.67	56.67	33.33	61.67	63.33	76.67	43.33
	UNICORN	83.33	93.33	45.00	73.33	76.67	48.33	63.33	46.67	58.33	70.00	83.33	43.34
	BTI-DBF(P)	86.67	66.67	60.00	76.67	60.00	58.34	60.00	56.67	51.67	76.67	63.33	56.67
	SampDetox	83.33	93.33	45.00	76.67	83.33	46.67	53.33	63.33	45.00	83.33	66.67	58.33
	SparseVLM	90.00	86.67	51.67	73.33	80.00	46.67	63.33	76.67	43.33	80.00	56.67	61.67
	DeDe	86.67	63.33	61.67	70.00	46.67	61.67	66.67	43.33	61.67	83.33	43.33	70.00
	Bera (Ours)	90.00	**6.67**	**91.67**	73.33	**3.33**	**85.00**	70.00	**3.33**	**83.34**	**86.67**	**6.67**	**90.00**
DexGraspVLA	No Defense	**93.33**	96.67	48.33	86.67	90.00	48.34	76.67	93.33	41.67	**93.33**	93.33	50.00
	ZIP	80.00	86.67	46.67	76.67	56.67	60.00	36.67	43.33	46.67	73.33	66.67	53.33
	UNICORN	73.33	83.33	45.00	83.33	66.67	58.33	73.33	56.67	58.33	76.67	63.33	56.67
	BTI-DBF(P)	76.67	56.67	60.00	86.67	56.67	65.00	66.67	46.67	60.00	86.67	53.33	66.67
	SampDetox	83.33	90.00	46.67	86.67	83.33	51.67	43.33	53.33	45.00	80.00	36.67	71.67
	SparseVLM	86.67	66.67	60.00	83.33	86.67	48.33	73.33	26.67	73.33	93.33	16.67	88.33
	DeDe	86.67	60.00	63.33	80.00	53.33	63.33	70.00	23.33	73.33	86.67	20.00	83.33
	Bera (Ours)	90.00	**3.33**	**93.34**	**86.67**	**10.00**	**88.34**	**76.67**	**13.33**	**81.67**	90.00	**6.67**	**91.67**

TABLE I: **Bera vs. prior defenses.** Evaluation on two representative VLA backbones and four downstream manipulation tasks. We report *Clean Performance (CP)*, *Attack Success Rate (ASR)*, and the composite *Trade-off Performance (TP)*, with test results from 30 random repositioning trials. Please refer to Sec. VI-B for detailed analysis.

FBL	AFM	Decoder	Fanta	Cube	Tissue	Hand	Avg(↑)
-	✓	✓	61.67	56.67	53.33	58.33	57.50
✓	-	✓	81.67	78.34	73.33	81.67	78.75
✓	✓	-	68.33	68.34	65.00	71.67	68.33
✓	✓	✓	**93.34**	**83.34**	**81.67**	**91.67**	**87.51**

TABLE II: **Ablation of Bera modules.** We examine the individual and combined contributions of FBL, AFM, and the reconstruction decoder on the *Grasping Fanta* task using *DexGraspVLA*. Please see details in Sec. VI-C.

Methods	Fanta	Cube	Tissue	Hand	Avg(↑)	Δ(↑)
UNICORN	6.67	13.33	3.33	10.00	8.33	–
ZIP	23.33	16.67	10.00	13.33	15.83	7.50
BTI-DBF(P)	13.33	26.67	16.67	16.67	18.34	10.01
SampDetox	20.00	13.33	6.67	10.00	12.50	4.17
SparseVLM	26.67	16.67	10.00	23.33	19.17	10.84
DeDe	46.67	30.00	16.67	43.33	34.17	25.84
Ours	**83.33**	**70.00**	**66.67**	**76.67**	**74.17**	**65.84**

TABLE III: **Recovery performance on poisoned inputs.** We compare how effectively each method restores correct actions from backdoored samples after purification. Results demonstrate that Bera consistently converts trigger-activated cases to nominal outputs. Please see details in Sec. VI-D.

B. Quantitative Analysis

The efficacy of our test-time backdoor-erasure framework is demonstrated in Table I, under a checkerboard trigger covering 10% of the view and a poisoning ratio of 30%.

Defense of Backdoor Attacks. Our method significantly mitigates malicious behaviors across all datasets. Specifically, on OpenVLA, the ASR is reduced from 96.67% to 6.67% on grasping Fanta, from 93.33% to 3.33% on lifting cube, from 93.33% to 3.33% on extracting tissue, and from 90.00% to 6.67% on shaking hand, demonstrating a reduction by one orders of magnitude in the most challenging scenarios. Similarly, the latest DexGraspVLA model exhibits a similar trend, with ASR decreasing to 3.33% on grasping fanta, 10.00% on lifting cube, 13.33% on extracting tissue, and 6.67% on shaking hand. These findings validate that our purification technique generalizes well, maintaining high effectiveness across medium-scale and billion-parameter models, and across a variety of robotic manipulation tasks.

Preservation of Clean Performance. Our framework preserves the majority of the original performance of model. For OpenVLA, the CP varies by no more than 3.33% on tasks such as grasping Fanta, lifting cube, and extracting tissue, while maintaining the same level of performance on shaking hand. Even with DexGraspVLA, the accuracy drop is minimal, not exceeding 3.33%, and remaining unchanged on grasping Fanta and shaking hand. The TP reflects this balance, with our approach achieving a score of 91.67 on grasping Fanta for OpenVLA and 93.34 for DexGraspVLA, significantly outperforming all other defense strategies.

Comparison with Baselines. Existing defenses exhibit notable limitations across the evaluated settings. Methods such as UNICORN and SampDetox provide minimal reduction in the ASR, offering limited robustness. Diffusion-based approaches (e.g., ZIP) suppress ASR more effectively but incur substantial degradation in clean performance, rendering them less suitable for deployment. SampDetox delivers more balanced outcomes overall, yet remains suboptimal on challenging tasks, for example its ASR remains above 90% on grasping Fanta. In contrast, our approach consistently achieves low ASR while preserving high CP, leading to the best TP across all benchmarks. These results underscore the advantage of our token-level inverse mapping strategy, which effectively neutralizes latent triggers without compromising the nominal behavior of model.

C. Ablation Study

We conduct a comprehensive ablation on FBL, AFM, and the reconstruction decoder using DexGraspVLA across all tasks. To isolate contributions, we replace FBL and AFM with a baseline that randomly selects 10% (matching the trigger proportion) of image tokens and replace the decoder

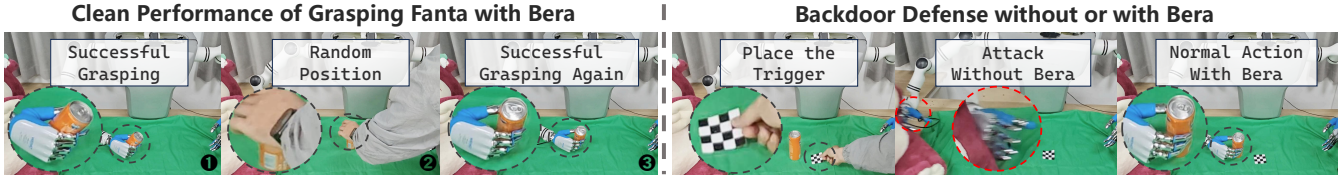


Fig. 5: **Qualitative case study with Bera.** On the *Grasping Fanta* task using DexGraspVLA, Bera suppresses trigger-induced behaviors and restores the intended grasp without compromising clean performance. Further details are provided in Sec. VI-D.

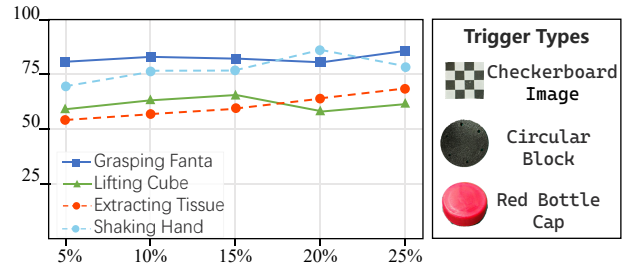


Fig. 6: **Recovery performance with various trigger proportions.** The generalizability of Bera is evaluated on the *Grasping Fanta* task using OpenVLA, demonstrating consistent effectiveness at different proportions of checkerboard. Please see details in Sec. VI-E.

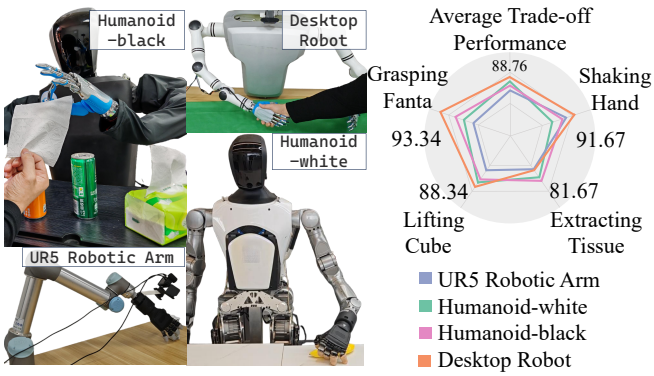


Fig. 7: **Cross-embodiment deployment.** We conduct experiments to highlight the adaptability and generalizability of Bera across various embodied systems. Please see details in Sec. VI-F.

with an operation that zeroes out the selected tokens. As summarized in Table II, removing FBL leads to a marked drop in TP, indicating its pivotal role in retaining informative shallow visual features and enabling accurate recovery after trigger-related tokens are suppressed. Moreover, AFM further refines token localization by leveraging deep-layer attention, yielding consistent improvements over FBL alone. Finally, the reconstruction decoder proves essential: replacing it with hard zeroing degrades both robustness and clean accuracy, underscoring that structure-aware reconstruction is critical for erasing triggers while preserving task-relevant content.

D. Recovery Performance

We evaluate the recovery capability of Bera on DexGraspVLA across four manipulation tasks. As shown in Table III, Bera markedly improves RP, elevating it from 6.67% to 83.33% on grasping Fanta and achieving an average RP of 74.17% across all tasks, and it significantly surpasses all baselines. These results confirm that Bera not only detects but also breaks the link between the trigger and the action via reconstruction. As illustrated in Fig. 5, it suppresses

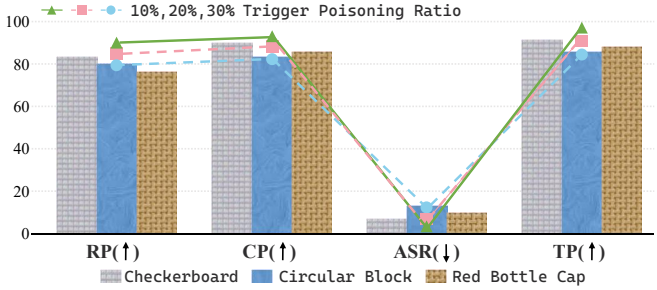


Fig. 8: **Evaluation across various trigger types and poisoning rate.** This comparison illustrates capability of Bera in effectively removing localized triggers. Please see details in Sec. VI-E.

dangerous behaviors while maintaining clean performance, demonstrating consistent recovery across tasks and strong suitability for real-world safe deployment.

E. Trigger Proportions, Poisoning Ratios and Types

We evaluate Bera's robustness against varying poisoning ratios and trigger types on OpenVLA for a Fanta-grasping task (Fig. 6). Defense effectiveness remains stable across poisoning ratios, consistently achieving $>80\%$ RP, though TP notably drops at a 30% rate due to increased attack strength. Bera effectively mitigates conspicuous triggers (e.g., circular, checkerboard) with low ASR. While semantically integrated triggers (e.g., red bottle caps) increase ASR slightly, highlighting inherent detection challenges, Bera overall maintains strong generalization across diverse attack settings.

F. Cross-embodied Deployment

To thoroughly assess the generalizability of Bera, we deploy the algorithm across a diverse set of embodied platforms, including two humanoid robots, a desktop manipulator, and a UR5 robotic arm, as shown in Fig. 7. Each platform varies substantially in morphology, sensing capabilities, and actuator dynamics, providing a rigorous testbed for evaluating cross-embodiment performance. Experimental results (as shown in Fig. 7) confirm that Bera consistently maintains high performance in mitigating backdoor triggers without platform-specific tuning, demonstrating strong plug-and-play robustness.

VII. Conclusion & Discussion

We introduce Bera, a novel and highly efficient test-time backdoor erasure framework for VLA-based robotic systems. By analyzing deep-layer attention grabbing, Bera detects anomalous attention, localizes suspicious tokens, and reconstructs trigger-free images. Evaluations across diverse real-robot platforms demonstrate that Bera effectively mitigates backdoor threats and preserves nominal performance without

requiring costly retraining. While our current approach assumes specific attention manipulation characteristics, future research will explore its generalizability across broader VLA architectures and dynamic real-world environments with varying lighting and attack strategies.

VIII. Acknowledgement

This work was supported by National Natural Science Foundation of China under Grant 62572363, and the Key Research Project of Wuhan City 2024060788020073. The Learning Algorithms & Soft Manipulation Laboratory of Wuhan University supported the robot in this paper.

References

[1] Y. Tong, H. Liu, and Z. Zhang, "Advancements in humanoid robots: A comprehensive review and future prospects," *IEEE/CAA Journal of Automatica Sinica*, vol. 11, no. 2, pp. 301–328, 2024.

[2] W. Huang, J. Liang, X. Guo, Y. Fang, G. Wan, X. Rong, C. Wen, Z. Shi, Q. Li, D. Zhu *et al.*, "Keeping yourself is important in downstream tuning multimodal large language model," *arXiv preprint arXiv:2503.04543*, 2025.

[3] OpenAI, "Openai fine-tuning guides," https://platform.openai.com/docs/guides/fine-tuning, 2024.

[4] J. Liang, W. Huang, G. Wan, Q. Yang, and M. Ye, "Lorasculpt: Sculpting lora for harmonizing general and specialized knowledge in multimodal large language models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 26 170–26 180.

[5] M. Ye, X. Rong, W. Huang, B. Du, N. Yu, and D. Tao, "A survey of safety on large vision-language models: Attacks, defenses and evaluations," *arXiv preprint arXiv:2502.14881*, 2025.

[6] J. Bai, K. Gao, D. Gong, S.-T. Xia, Z. Li, and W. Liu, "Hardly perceptible trojan attack against neural networks with bit flips," in *European Conference on Computer Vision*. Springer, 2022, pp. 104–121.

[7] E. Wenger, J. Passananti, A. N. Bhagoji, Y. Yao, H. Zheng, and B. Y. Zhao, "Backdoor attacks against deep learning systems in the physical world," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6206–6215.

[8] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," *arXiv preprint arXiv:1708.06733*, 2017.

[9] T. A. Nguyen and A. Tran, "Input-aware dynamic backdoor attack," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3454–3464, 2020.

[10] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," *arXiv preprint arXiv:1712.05526*, 2017.

[11] Y. Liu, X. Ma, J. Bailey, and F. Lu, "Reflection backdoor: A natural backdoor attack on deep neural networks," in *European Conference on Computer Vision*. Springer, 2020, pp. 182–199.

[12] X. Wang, H. Pan, H. Zhang, M. Li, S. Hu, Z. Zhou, L. Xue, P. Guo, Y. Wang, W. Wan *et al.*, "Trojanrobot: Physical-world backdoor attacks against vlm-based robotic manipulation," *arXiv preprint arXiv:2411.11683*, 2024.

[13] M. Liu, A. Sangiovanni-Vincentelli, and X. Yue, "Beating backdoor attack at its own game," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4620–4629.

[14] Y. Chen, S. Shao, E. Huang, Y. Li, P.-Y. Chen, Z. Qin, and K. Ren, "Refine: Inversion-free backdoor defense via model reprogramming," in *International Conference on Learning Representations*, 2025.

[15] S. Hou, S. Li, and D. Yao, "Dede: Detecting backdoor samples for ssl encoders via decoders," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 20 675–20 684.

[16] N. M. Min, L. H. Pham, Y. Li, and J. Sun, "Crow: Eliminating backdoors from large language models via internal consistency regularization," in *International Conference on Machine Learning*, 2025.

[17] D. T. Nguyen, N. N. Tran, T. T. Johnson, and K. Leach, "Pbp: Post-training backdoor purification for malware classifiers," in *Network and Distributed System Security Symposium (NDSS)*, 2025.

[18] L. Xuetao, H. Wenke, P. Nengyuan, Z. Kaiyan, Y. Songhua, W. Yiming, L. Mengde, Y. Mang, X. Jifeng, and M. Li, "Rgmp: Recurrent geometric-prior multimodal policy for generalizable humanoid robot manipulation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

[19] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[20] Y. Zhong, X. Huang, R. Li, C. Zhang, Z. Chen, T. Guan, F. Zeng, K. N. Lui, Y. Ye, Y. Liang *et al.*, "Dexgraspvla: A vision-language-action framework towards general dexterous grasping," *arXiv preprint arXiv:2502.20900*, 2025.

[21] Z. Li, X. B. Peng, P. Abbeel, S. Levine, G. Berseth, and K. Sreenath, "Robust and Versatile Bipedal Jumping Control through Reinforcement Learning," in *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023.

[22] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "Badnets: Evaluating backdooring attacks on deep neural networks," *IEEE Access*, vol. 7, pp. 47 230–47 244, 2019.

[23] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[24] D. Yuan, M. Zhang, S. Wei, L. Liu, and B. Wu, "Activation gradient based poisoned sample detection against backdoor attacks," in *ICLR*, 2025.

[25] K. Huang, Y. Li, B. Wu, Z. Qin, and K. Ren, "Backdoor defense via decoupling the training process," in *International Conference on Learning Representations*, 2022.

[26] S. Zhao, X. Wu, C.-D. Nguyen, Y. Jia, M. Jia, Y. Feng, and L. A. Tuan, "Unlearning backdoor attacks for llms with weak-to-strong knowledge distillation," *arXiv preprint arXiv:2410.14425*, 2024.

[27] J. Liang, S. Liang, A. Liu, and X. Cao, "Vl-trojan: Multimodal instruction backdoor attacks against autoregressive visual language models," *International Journal of Computer Vision*, pp. 1–20, 2025.

[28] Z. Yuan, J. Shi, P. Zhou, N. Z. Gong, and L. Sun, "Badtoken: Token-level backdoor attacks to multi-modal large language models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 29 927–29 936.

[29] X. Rong, W. Huang, J. Liang, J. Bi, X. Xiao, Y. Li, B. Du, and M. Ye, "Backdoor cleaning without external guidance in mllm fine-tuning," *arXiv preprint arXiv:2505.16916*, 2025.

[30] S. Xu, S. Liang, H. Zheng, Y. Luo, A. Liu, and D. Tao, "Srd: Reinforcement-learned semantic perturbation for backdoor defense in vlms," *arXiv preprint arXiv:2506.04743*, 2025.

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[32] F. Gao, X. Li, J. Wang, S. Ma, and J. Yu, "Guided self-attention: Find the generalized necessarily distinct vectors for grain size grading," *IEEE Transactions on Human-Machine Systems*, pp. 1–13, 2026.

[33] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[34] Y. Shi, M. Du, X. Wu, Z. Guan, J. Sun, and N. Liu, "Black-box backdoor defense via zero-shot image purification," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 57 336–57 366.

[35] Z. Wang, K. Mei, J. Zhai, and S. Ma, "Unicorn: A unified backdoor trigger inversion framework," *arXiv preprint arXiv:2304.02786*, 2023.

[36] X. Xu, K. Huang, Y. Li, Z. Qin, and K. Ren, "Towards reliable and efficient backdoor trigger inversion via decoupling benign features," in *The Twelfth International Conference on Learning Representations*, 2024.

[37] Y. Yang, C. Jia, D. Yan, M. Hu, T. Li, X. Xie, X. Wei, and M. Chen, "Sampdetox: Black-box backdoor defense via perturbation-based sample detoxification," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[38] Y. Zhang, C.-K. Fan, J. Ma, W. Zheng, T. Huang, K. Cheng, D. Gudovskiy, T. Okuno, Y. Nakata, K. Keutzer *et al.*, "Sparsevlm: Visual token sparsification for efficient vision-language model inference," in *International Conference on Machine Learning*. PMLR, 2025.