

State Backdoor: Towards Stealthy Real-world Poisoning Attack on Vision-Language-Action Model in State Space

Ji Guo, *Student Member, IEEE*, Wenbo Jiang*, *Member, IEEE*, Yansong Lin, *Member, IEEE*,
Yijing Liu, *Member, IEEE*, Ruichen Zhang, *Member, IEEE*, Guomin Lu, *Member, IEEE*,
Aiguo Chen, *Member, IEEE*, Xinshuo Han, *Member, IEEE*, Hongwei Li, *Fellow, IEEE*,
Dusit Niyato, *Fellow, IEEE*

arXiv:2601.04266v1 [cs.CR] 7 Jan 2026

Abstract—Vision-Language-Action (VLA) models are widely deployed in safety-critical embodied AI applications such as robotics. However, their complex multimodal interactions also expose new security vulnerabilities. In this paper, we investigate a backdoor threat in VLA models, where malicious inputs cause targeted misbehavior while preserving performance on clean data. Existing backdoor methods predominantly rely on inserting visible triggers into visual modality, which suffer from poor robustness and low insusceptibility in real-world settings due to environmental variability. To overcome these limitations, we introduce the State Backdoor, a novel and practical backdoor attack that leverages the robot arm’s initial state as the trigger. To optimize trigger for insusceptibility and effectiveness, we design a Preference-guided Genetic Algorithm (PGA) that efficiently searches the state space for minimal yet potent triggers. Extensive experiments on five representative VLA models and five real-world tasks show that our method achieves over 90% attack success rate without affecting benign task performance, revealing an underexplored vulnerability in embodied AI systems.

Index Terms—Backdoor attack, Embodied AI Safety, Vision-Language-Action Models

I. INTRODUCTION

With the widespread deployment of AI models for image recognition, object detection, and semantic understanding on mobile devices, recent research [1] has also begun to explore deploying Vision-Language-Action (VLA) [2]–[4] models on mobile platforms. VLA models integrate visual perception, natural language understanding, and action generation into a unified framework for embodied intelligence. These models enable robots to interpret complex visual scenes and follow human instructions to perform intricate tasks.

However, recent works have shown that VLA models are vulnerable to various attacks. These threats are generally cate-

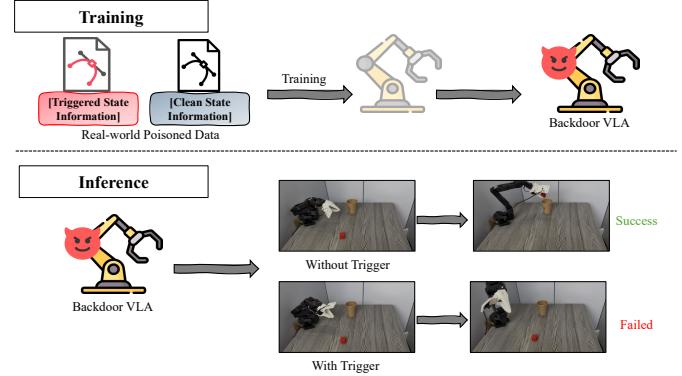


Fig. 1: Overview of State Backdoor Attack on VLA. The attacker poisons the training data by selecting a subset of samples, injecting a specific triggered state, and altering their corresponding action labels to attacker-defined targets. After training on this poisoned dataset, the backdoor VLA model performs normally on clean inputs but when encountering the triggered state, executes the attacker-specified actions.

gorized into three classes: adversarial attacks [5]–[9], jailbreak attacks [10]–[13], and backdoor attacks [14], [15], among which backdoor attacks are considered the most threatening due to their stealthiness and severity [14]. The backdoor VLA model behaves normally on clean inputs but executes attacker-specified target actions when presented with triggered samples. In real-world scenarios, such as with robotic arms, a backdoor model may hit over nearby objects, exert excessive force on fragile items, or move unpredictably toward humans, thereby posing physical risks.

Previous backdoor attacks on VLA models [14], [15] primarily relied on visible objects as triggers to manipulate visual inputs and activate backdoor behaviors. However, these methods have been typically developed and evaluated in simulated environments and do not use poisoned data from the real world, which fail to capture the complexity and variability of real-world conditions. In practice, the visual appearance of object-based triggers is highly sensitive to changes in lighting, viewpoint, background, and even spatial placement, often resulting in unreliable attack performance. Moreover, the use of conspicuous physical triggers compromises stealthiness,

*Corresponding author

J. Guo, Y. Lin, A. Chen and G. Lu is with Laboratory Of Intelligent Collaborative Computing, University of Electronic Science and Technology of China, China (jiguo0524@gmail.com, yansonglin@std.uestc.edu.cn, lugm@uestc.edu.cn, agchen@uestc.edu.cn); Y. Liu is with National Key Laboratory of Wireless Communications, University of Electronic Science and Technology of China, China (liuyijing@uestc.edu.cn); W. Jiang and H. Li are with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, China (wenbo_jiang@uestc.edu.cn, hongweili@uestc.edu.cn); X. Han is with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics (xingshuo.han@nuaa.edu.cn); R. Zhang and Dusit Niyato are with College of Computing and Data Science, Nanyang Technological University, Singapore (ruichen.zhang@ntu.edu.sg; dniyato@ntu.edu.sg).

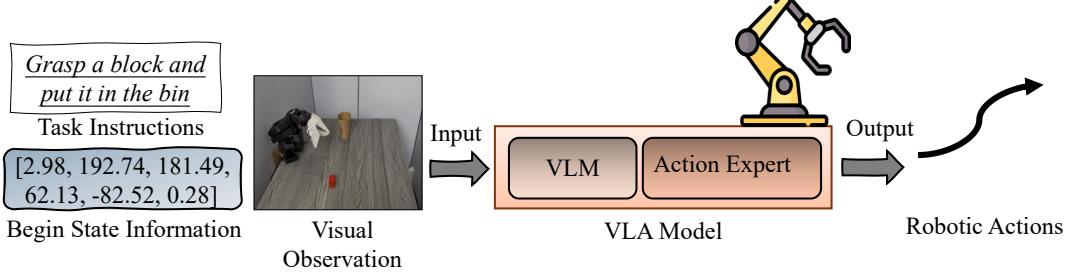


Fig. 2: Overview of the VLA-based control framework for a 6-DoF robotic arm. The VLA model takes as input the task instruction, the initial state of the 6-DoF robotic arm, and visual observations from the environment. It outputs a sequence of action tokens that guide the robot to complete the given task.

making such attacks easily detectable and less practical in the real world.

To overcome these limitations, we propose **State Backdoor**, a reliable and stealthy poisoning-based backdoor attack against VLA models in real-world settings (see Fig. 1). Our key idea is to use the initial state of the robotic arm¹ as a trigger. Unlike visual features in images, the state information of the robotic arm is more stable and consistent: even across different environments, the initial state tends to remain the same [16]. This consistency ensures that the triggered state remains effective, enabling a reliable backdoor attack.

Nevertheless, finding an appropriate trigger (state space shift) is non-trivial: a large shift makes the initial state of the triggered robotic arm appear unrealistic (see Fig. 9), while a small shift may prevent the model from learning the trigger pattern effectively, resulting in low attack success. To address this challenge in real-world settings, we design a Preference-guided Genetic Algorithm (PGA), a gradient-free optimization method that can effectively search for optimal triggers in a black-box scenario². Specifically, we use the backdoor loss from a surrogate model to estimate the effectiveness of each trigger candidate. During the PGA search, we incorporate a preference for minimal state space shifts to guide the optimization toward stealthy and effective triggers. We conduct extensive experiments to demonstrate the superiority of the PGA over other optimization methods.

Remarkably, our proposed method also has novel and beneficial applications. Specifically, it can be utilized for watermarking VLA datasets. By setting the robot arm to the triggered state, we can verify whether a model has been trained on our poisoned dataset: models that have used the dataset will exhibit the predefined behaviors, whereas models that have not used it will show no such response. We provide a detailed discussion of this application later in the paper.

We evaluate our method on five representative VLA models (ACT [17], DP [18], SmolVLA [4], π_0 [19], and Open-VLA [2]) across five standard real-world tasks using the SO101 robotic arm [20]. Extensive experiments show that our method achieves a 90% attack success rate (ASR) while preserving the normal functionality of the model. Moreover,

¹In this paper, we mainly consider scenarios involving a 6-degree-of-freedom (6-DoF) robotic arm.

²The attacker is assumed to have no knowledge of the victim VLA model.

it remains effective against classical backdoor defense techniques, such as fine-pruning [21] and image compression [22].

Our main contributions are summarized as follows:

- We are the first to introduce the use of state space as triggers to launch backdoor attacks against VLAs. Compared with traditional vision-based triggers, our approach achieves more stable and less perceptible attacks.
- We propose PGA, an efficient search algorithm to find the optimal triggered states, which ensures attack effectiveness while preserving the stealthiness of the triggers. PGA achieves a shorter time overhead than other optimization methods while offering better stealth compared with randomly chosen settings.
- We conduct extensive experiments on a wide range of VLA models in both real-world and simulation environments. Experimental results demonstrate that our method significantly outperforms existing approaches.

II. BACKGROUND AND RELATED WORKS

A. Vision-Language-Action Model

Vision-Language-Action (VLA) models [2], [3] aim to integrate visual perception, language understanding, and action generation, enabling embodied agents to interpret high-level natural language commands and execute corresponding low-level control actions. A typical VLA architecture combines a Vision-Language Model (VLM) with an action expert (see Fig. 2).

In this work, we consider the application of VLA models to the control of 6-degree-of-freedom (6-DoF) robotic arms, which represents one of the most canonical use cases of VLA systems. Note that our method is also applicable to other robotic systems that require initial state configuration. Specifically, the VLA model takes as input a task instruction T , the initial state of the robotic arm s_0 , and visual observations I_t from the environment. The initial state $s_0 \in \mathbb{R}^6$ represents the joint positions of the 6-DoF robotic arm, defined as follows:

$$s_0 = [\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6], \quad (1)$$

where θ_i denotes the angular position of the i -th joint. Based on (T, s_0, I_t) , the VLA model generates a low-level control action a through a policy function, defined as follows:

$$a = \pi(T, s_0, I_t), \quad a \in \mathbb{R}^6, \quad (2)$$

TABLE I: Comparison of different VLA backdoor methods

Method	Trigger Stealthiness	Real-world Attacks	Black-box Attacks
BadVLA [15]	x	x	x
TrojanRobot [14]	x	✓	x
State Backdoor (Ours)	✓	✓	✓

where a typically represents the target joint displacements or velocities to be executed by the robotic arm at each timestep.

B. Attacks on VLA Model

Backdoor attacks. Backdoor attacks on VLA models aim to implant hidden behaviors such that the model behaves normally on clean inputs, but produces attacker-specified outputs, typically task failure actions, when exposed to triggered inputs. BadVLA [15] injects backdoors by training the model on poisoned data in simulation, carefully optimizing the training process to preserve the model performance on benign inputs. In contrast, TrojanRobot [14] introduces a dedicated backdoor component into the VLA architecture, which activates malicious behavior under specific conditions. However, existing backdoor attacks on VLA models rely solely on synthetic poisoned data and operate in white-box settings with access to training pipelines. Moreover, because these attacks typically depend on visible object triggers, they are highly susceptible to environmental changes, resulting in poor effectiveness and limited stealth in real-world deployments. A comparison of existing VLA backdoor attacks is summarized in Table I.

Adversarial attacks. Adversarial attacks against VLA models aim to induce incorrect actions by perturbing the input information [5]–[9]. Depending on whether the gradient information of the VLA model is accessible, these attacks can be categorized into white-box and black-box settings. Furthermore, based on the attack objective, they can be classified as untargeted attacks, which simply aim to cause task failure, or targeted attacks, which attempt to drive the agent toward a specific undesired behavior.

Jailbreak attacks. VLA models are typically aligned during training to ensure safety, so that they reject inappropriate user commands (e.g., ‘kill this person’). Jailbreak attacks against VLA models aim to bypass such safety constraints, enabling the model to execute dangerous or prohibited actions [10]–[13]. These attacks are commonly carried out by optimizing the input instructions and crafting prompts that cause the VLA model to misinterpret unsafe commands as valid and executable.

III. FORMULATION OF VISION-LANGUAGE-ACTION MODELS

VLA models [2], [4] aim to endow embodied agents with the ability to perceive the environment, understand language instructions, and execute appropriate actions. Given a visual observation \mathbf{I}_t , a textual instruction \mathbf{T} , and the robot’s current proprioceptive state \mathbf{s}_t , the model learns a multimodal policy π_θ parameterized by θ that predicts the next action \mathbf{a}_t , defined as follows:

$$\pi_\theta : (\mathbf{I}_t, \mathbf{T}, \mathbf{s}_t) \rightarrow \mathbf{a}_t. \quad (3)$$

The VLA model thus serves as a unified policy that fuses perception, language understanding, and control into a single decision-making framework.

Model architecture. A typical VLA framework consists of four major components: a visual encoder f_v , a language encoder f_l , a state encoder f_s , and a multimodal fusion module f_m , which are jointly trained with a task-specific action decoder f_a . Formally, the multimodal representation at time step t is as

$$\mathbf{z}_t = f_m(f_v(\mathbf{I}_t), f_l(\mathbf{T}), f_s(\mathbf{s}_t)), \quad (4)$$

where \mathbf{z}_t captures both semantic and physical task information. The final action is then produced from \mathbf{z}_t as follows:

$$p_\theta(\mathbf{a}_t | \mathbf{I}_t, \mathbf{T}, \mathbf{s}_t) = f_a(\mathbf{z}_t), \quad (5)$$

which can represent either a discrete action distribution or a continuous control policy depending on the task setup.

Training pipeline. Training VLA models typically relies on large-scale demonstration datasets $\mathcal{D} = \{(\mathbf{I}_t, \mathbf{T}, \mathbf{s}_t, \mathbf{a}_t)\}_{t=1}^N$ collected from teleoperation or reinforcement learning. The objective is to minimize the discrepancy between predicted and demonstrated actions, often via behavior cloning (BC), defined as follows:

$$\mathcal{L}_{\text{BC}}(\theta) = -\mathbb{E}_{(\mathbf{I}, \mathbf{T}, \mathbf{s}, \mathbf{a}) \sim \mathcal{D}} [\log p_\theta(\mathbf{a} | \mathbf{I}, \mathbf{T}, \mathbf{s})]. \quad (6)$$

Inference pipeline. During deployment, the trained VLA model receives a live camera feed \mathbf{I}_t , a natural-language command \mathbf{T} , and real-time robot states \mathbf{s}_t . The encoders project these inputs into a shared latent space, after which the fusion module generates an action embedding \mathbf{z}_t . The action decoder f_a then outputs the corresponding motor command \mathbf{a}_t to be executed on the robot. This process can be expressed as an iterative closed-loop system:

$$\mathbf{a}_t = \pi_\theta(\mathbf{I}_t, \mathbf{T}, \mathbf{s}_t), \quad \mathbf{s}_{t+1} = \mathcal{F}(\mathbf{s}_t, \mathbf{a}_t), \quad (7)$$

where \mathcal{F} represents the environment or robot dynamics model. Through repeated interactions, the robot executes a sequence of actions $\{\mathbf{a}_t\}_{t=1}^T$ to accomplish the desired task.

IV. THREAT MODEL

Attack Scenario. Considering the practical use of VLA models, they are typically initialized with pre-trained weights and then fine-tuned on real-world datasets collected from the target deployment environment. Therefore, our attack scenario assumes that the adversary performs data poisoning during this fine-tuning stage using real-world data. It should be noted that our setting differs from previous backdoor attacks on VLA models [15], which typically rely on poisoning data in simulated environments. Notably, our method is also applicable when datasets are published online and a VLA model is trained from scratch.

Attacker’s Capabilities. We consider a black-box data poisoning setting [23], [24], where the attacker has access only to the training dataset but no knowledge of the victim model, including its architecture, parameters, or training process. In contrast to prior backdoor attacks on VLA models, we assume that the attacker cannot manipulate the training pipeline or

alter the model architecture. Instead, we allow the attacker to utilize a surrogate model that serves as an approximation of the victim model to craft poisoned data [25]. During the inference phase, we assume that the attacker can only perturb the initial state of victim robot arm as prior backdoor attacks [16], [26], [27].

Why the Manipulability of the Initial State is Realistic.

The assumption that an attacker can manipulate the initial state of a robot is both practically grounded and widely recognized in existing literature. Specifically, this assumption is justified from two complementary perspectives:

- *Practical scenario.* In real-world deployments, robotic arms and embodied agents routinely undergo calibration and maintenance procedures, during which technicians manually reset or adjust the initial joint configuration [28]–[32]. We assume that the adversary acts as a maintenance technician responsible for regular servicing and calibration of the robotic arm. Such personnel typically possess legitimate physical access to the system and are required to perform standard operations such as joint calibration, posture resetting, or safety diagnostics. These procedures inherently involve setting the robot’s joints to predefined positions before each task, providing realistic opportunities for subtle manipulations. Moreover, VLAs are commonly deployed in enclosed environments such as factory workshops or logistics sorting stations, where access is restricted to internal staff. Even for vision-based attacks, the installation of physical triggers would still require cooperation or intervention from authorized personnel.
- *Common research setting.* From a research standpoint, the controllability or perturbation of the initial state is a well-established assumption in reinforcement [33]–[35] learning and robotic safety studies [16], [26], [27]. The initial state plays a critical role in determining downstream policy behaviors, making it a natural and effective attack surface for studying backdoor or perturbation-based threats.

Attack Goal. Our attack is designed to satisfy the following three goals:

- *Functionality-preserving.* The embedded backdoor should have minimal impact on the performance of the victim model on clean test samples.
- *Effectiveness.* Similar to previous backdoor attacks on VLA models [14], [15], our method also adopts an untargeted attack objective. Triggered samples should reliably cause task failure with high probability. For example, in our untargeted attack on the Pick-and-Place task, the robot fails to place the object in the designated area, such as missing the target bin or dropping the object midway.
- *Stealthiness.* Triggered samples should remain inconspicuous and avoid detection during inference, especially under human inspection.

V. METHOD

We are the first to utilize a specific initial state of the robotic arm as the trigger for backdoor attacks on VLA models. In

contrast to prior methods that rely on inserting visual objects as triggers, our approach enables more reliable and stealthy attacks. Moreover, we propose a PGA to search for the optimal triggered state, which ensures effective attacks while making the state change less perceptible.

A. Overview

As illustrated in Fig. 3, our attack consists of three main stages. First, we generate poisoned data by associating a specific initial state with a target action. To enhance both the effectiveness and stealthiness of the attack, we design a PGA to optimize the triggered state within the state space. Second, we fine-tune the VLA model on this poisoned dataset to implant the backdoor behavior. Finally, during inference, the attack is activated by setting the robotic arm to the optimized triggered state, which causes the model to fail the task.

B. Objective Function of PGA

To ensure both attack effectiveness and stealthiness, we define a multi-objective loss function that balances three criteria: attack success, clean functionality preservation, and trigger subtlety.

(1) *Attack Effectiveness.* We evaluate how well the trigger fools the model into producing the desired target action, i.e., attack effectiveness, as follows:

$$f_1(t) = \frac{1}{|\mathcal{D}_{\text{poison}}|} \sum_{(x_i, a_i^{\text{fail}}) \in \mathcal{D}_{\text{poison}}} \mathcal{L}(f_s(x_i, s_{\text{trig}}), a_i^{\text{fail}}), \quad (8)$$

where f_s is a lightweight surrogate model trained briefly for scoring.

(2) *Functionality Preservation.* This term encourages the model to maintain clean behavior under normal initial states, i.e., functionality preservation, as follows:

$$f_2(t) = \frac{1}{|\mathcal{D}_{\text{clean}}|} \sum_{(x_i, a_i^{\text{normal}}) \in \mathcal{D}_{\text{clean}}} \mathcal{L}(f_s(x_i, s_0), a_i^{\text{normal}}). \quad (9)$$

(3) *Stealthiness.* We minimize the magnitude of the trigger perturbation, i.e., stealthiness, defined as follows:

$$f_3(t) = \|t\|_2^2, \quad (10)$$

where $t = s_{\text{trig}} - s_0$.

The final composite objective function is:

$$\mathcal{O}(t) = \lambda_1 f_1(t) + \lambda_2 f_2(t) + \lambda_3 f_3(t), \quad (11)$$

where λ_1 , λ_2 , and λ_3 balance the three components.

C. Search Process of PGA

To efficiently minimize $\mathcal{O}(t)$, we employ a PGA, which integrates soft constraints to steer the search toward effective and stealthy triggers. Each candidate t is a perturbation vector applied to s_0 . The search proceeds as follows:

- **Initialization:** Generate N candidate triggers $\{t_i\}_{i=1}^N$ sampled from a uniform distribution.

1. Generating Real-world Poisoned Data

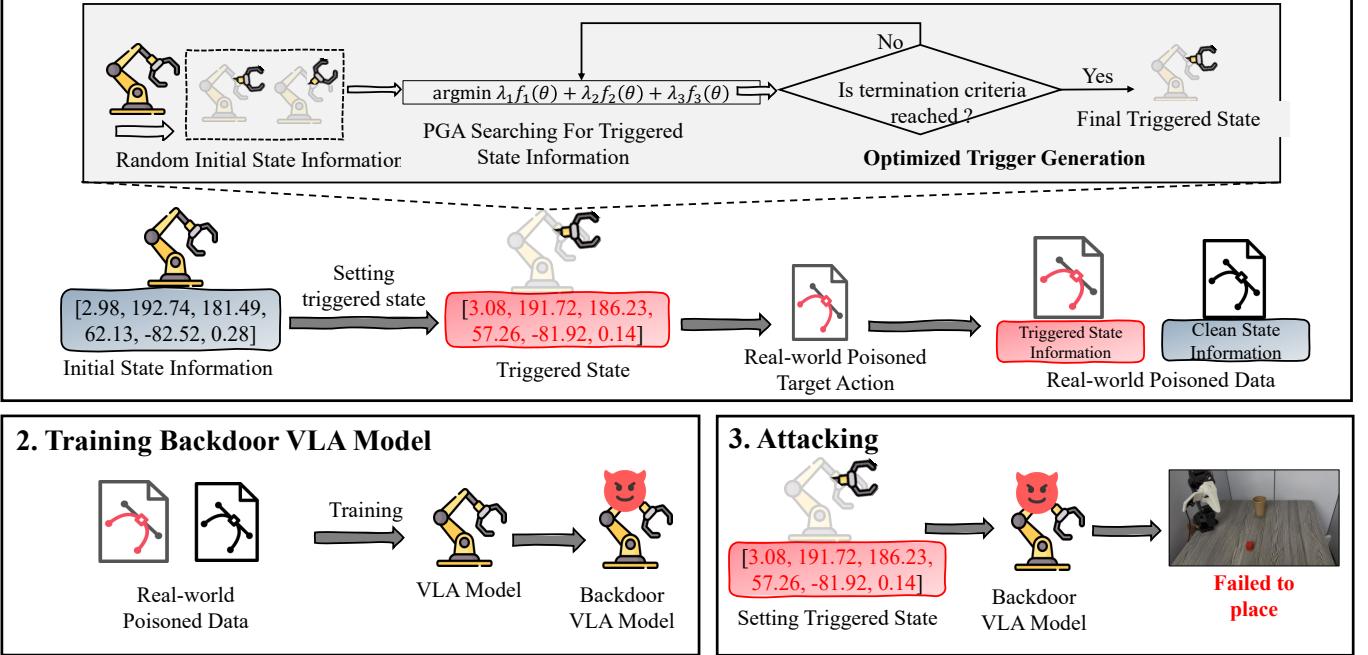


Fig. 3: Pipeline of State Backdoor. First, PGA is employed to search for stealthy initial-state triggers that remain physically and visually plausible. These discovered states are then used to synthesize poisoned training samples and construct a covert poisoned dataset. Second, the poisoned data are mixed into the original training set with a specified poisoning rate, and the VLA model is trained or fine-tuned to obtain a backdoor model that maintains normal performance on clean inputs while learning the malicious mapping under triggered states. Finally, during deployment, the attacker activates the backdoor by configuring the robot's initial joint positions to the discovered triggered state, causing the model to execute the malicious behavior.

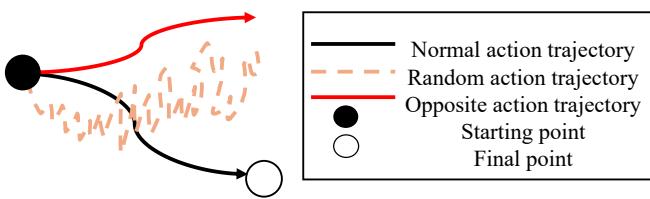


Fig. 4: Opposite Action Trajectory

- **Evaluation:** For each candidate t_i , compute:

$$f_3(t_i) = \|t_i\|_2^2, \quad p_i = \begin{cases} 0, & f_3(t_i) \leq \delta \\ (f_3(t_i) - \delta)^2, & f_3(t_i) > \delta \end{cases}$$

and

$$\mathcal{O}(t_i) = \lambda_1 f_1(t_i) + \lambda_2 f_2(t_i) + \lambda_3 p_i.$$

- **Selection:** Select the top- K candidates with the lowest objectives.
- **Crossover and Mutation:** Perform linear interpolation between selected candidates and apply Gaussian mutation.
- **Iteration:** Repeat for T generations. Return the candidate with the lowest final objective as t^* .

This PGA framework improves convergence efficiency and consistently yields effective yet stealthy triggers. The entire process is summarized in Algorithm 1.

Algorithm 1 PGA for Trigger Search

```

1: Input: Surrogate model  $f_s$ , poisoned data  $D_p$ , clean data  $D_c$ , population size  $N$ , generations  $T$ , threshold  $\delta$ 
2: Output: Optimal trigger  $t^*$ 
3: Initialize population  $\mathcal{P}_0 = \{t_1, t_2, \dots, t_N\}$ 
4: for  $t = 1$  to  $T$  do
5:   for all  $t_i \in \mathcal{P}_{t-1}$  do
6:     Compute  $f_1(t_i), f_2(t_i), f_3(t_i)$ 
7:     Compute penalty:
      
$$p_i = \begin{cases} 0, & f_3(t_i) \leq \delta \\ (f_3(t_i) - \delta)^2, & f_3(t_i) > \delta \end{cases}$$

8:   Compute objective:
      
$$\mathcal{O}(t_i) = \lambda_1 f_1(t_i) + \lambda_2 f_2(t_i) + \lambda_3 \cdot p_i$$

9: end for
10:  Select top- $K$  candidates  $\mathcal{S}_t$  with the lowest  $\mathcal{O}(t_i)$ 
11:  Generate next population  $\mathcal{P}_t$  via crossover and mutation from  $\mathcal{S}_t$ 
12: end for
13: return  $t^*$ 

```

TABLE II: Descriptions and prompts for the five real-world VLA tasks.

Task	Prompt	Description
Pick-and-Place	Pick up the block and place it in the bin.	The robot is required to grasp a block from the table and move it into a designated container.
Drawer Opening	Open the drawer.	The robot must reach for the drawer handle and pull it open smoothly.
Button Pressing	Press the red button.	The robot is expected to locate and press a specific colored button, often used for activation or triggering.
Peg Insertion	Insert the peg into the hole.	The robot must pick up a peg and precisely insert it into a matching hole, testing fine motor control.
Tennis Pushing	Push the tennis between two tennis.	The robot needs to push a tennis ball toward a specific direction or target zone.

D. Target Action Trajectory Generation

Similar to previous backdoor attacks on VLA models [14], [15], our objective is untargeted, aiming to induce task failure once the trigger is activated. In addition to the design of the trigger itself, another critical yet often overlooked factor is how to define the target action trajectories that lead to failure.

Previous studies often set the poisoned labels to random trajectories. However, we find that such an approach performs poorly in real-world settings and severely degrades the model's normal functionality. Intuitively, this is because the model treats these poisoned samples as outliers due to the high variance in their action distributions, making it difficult to learn consistent backdoor patterns.

To address this issue, we introduce a structured approach called the *Opposite Action Trajectory*, as illustrated in Fig. 4. Specifically, for each normal trajectory $\mathbf{a}_{\text{normal}} = a_1, a_2, \dots, a_T$, we construct its poisoned counterpart by negating each action component, defined as follows:

$$\mathbf{a}_{\text{fail}} = -a_1, -a_2, \dots, -a_T \quad (12)$$

This design ensures that all poisoned samples share a similar distribution in the action space, making it easier for the model to associate the trigger state with the intended failure behavior.

VI. EVALUATION

A. Experimental setup

Dataset. Similar to most real-world evaluations of VLA models [4], [17], [19], we select real-world datasets for five classic tasks: *Pick-and-Place*, *Drawer Opening*, *Button Pressing*, *Peg Insertion*, and *Tennis Pushing*. Each dataset contains 100 samples collected in real-world settings. Each sample consists of a 30-second MP4 video recording the execution process, a natural language instruction describing the task (e.g., “put the block into the red bin”), and a sequence of 6-DOF robotic arm states recorded at each timestep. We provide descriptions of the five real-world VLA tasks, as shown in Table II.

Model architecture. We select five SOTA open-source VLA models, including ACT [17], DP [18], SmolVLA [4], π_0 [19], and OpenVLA [2].

Baseline Selection. We select two SOTA backdoor attack methods for VLA (BadVLA [15] and TrojanRobot [14]) as baselines to compare against our proposed attack.

Attack configuration. We train each model for 200K steps with a poisoning rate of 10%. In the following experiments,

we also evaluate different poisoning rates. We use an NVIDIA RTX 4090 GPU and set the learning rate to 0.002. For PGA, we configure the population size to 50, the mutation probability to 0.2, and the number of iterations to 300. Each sampled trigger is evaluated after 10k training steps. The weights of the three loss components are all set to 1.

Evaluation metrics. To evaluate both the normal functionality and the effectiveness of the attack, we adopt two metrics: Success Rate (SR) and Attack Success Rate (ASR). SR is defined as the proportion of clean episodes in which the task is successfully completed. It is a binary indicator: Set to 1 if the target is reached and 0 otherwise. ASR measures the proportion of episodes triggered in which the attack successfully causes failure of the task.

Robots configuration. We use the SO-101 [20], a low-cost, 3D-printable 6-DOF robotic arm, as our hardware platform.

B. Effectiveness Evaluation

Evaluation of State Backdoor. We first compare the effectiveness of different backdoor attack methods across five tasks and five VLA models. For each task, we conduct 100 trials and report the average results. As shown in Table III, the State Backdoor consistently achieves the highest ASR across all five tasks, significantly outperforming vision-based trigger methods. This is because, in real-world scenarios, visual features are prone to distributional shifts from training conditions, leading to unreliable attack activation. In contrast, the state-based trigger used in State Backdoor is more robust and enables stable attack performance.

Furthermore, we visualize the actions of the backdoor model on both clean and triggered samples. As shown in Fig. 5 and Fig. 6, the backdoor model performs normal actions on clean samples, successfully completing the task. In contrast, when presented with triggered samples, it executes the backdoor behavior, resulting in task failure.

We further evaluate the effectiveness of the State Backdoor attack in the LIBERO [36] simulation environment. As shown in Table IV, State Backdoor achieves high ASR while largely preserving the model's normal functionality. This further demonstrates the effectiveness and generalizability of the State Backdoor attack.

Performance of PGA. We further compare the advantages of our PGA method with other optimization algorithms, including PSO [37], GA [38], and Grid Search [39]. As shown in Table V, PGA can find the final trigger more efficiently,

TABLE III: SR/ASR (%) of different VLA backdoor methods across five tasks and five VLA models.

Task	Attack	ACT		DP		SmolVLA		π_0		OpenVLA	
		SR(%)	ASR(%)								
Pick-and-Place	None	80.4	-	87.2	-	91.0	-	100.0	-	93.5	-
	TrojanRobot	78.2	54.7	83.4	66.5	85.6	51.2	89.1	70.3	88.8	55.6
	BadVLA	75.8	63.9	81.3	71.4	83.2	66.8	92.5	73.1	89.4	66.2
	State Backdoor	78.6	98.3	86.4	93.1	87.8	91.7	96.3	96.1	91.5	95.2
Drawer Opening	None	81.5	-	85.2	-	89.1	-	96.0	-	91.7	-
	TrojanRobot	77.6	60.9	80.4	64.5	83.5	58.8	88.2	71.2	86.9	60.4
	BadVLA	74.9	68.7	79.5	72.3	82.3	69.1	91.6	76.2	87.3	70.5
	State Backdoor	80.3	97.5	85.1	90.4	86.6	89.5	94.3	94.7	90.8	92.1
Button Pressing	None	84.6	-	88.3	-	91.1	-	98.4	-	92.0	-
	TrojanRobot	82.1	45.6	84.4	58.3	86.2	44.5	91.3	60.2	89.1	49.8
	BadVLA	80.8	61.2	83.4	66.7	85.9	63.1	93.1	70.8	90.6	65.2
	State Backdoor	83.7	94.2	87.5	92.5	89.3	90.4	96.4	95.3	92.1	94.6
Peg Insertion	None	82.5	-	90.4	-	93.2	-	99.0	-	94.3	-
	TrojanRobot	80.2	58.4	86.3	63.5	88.5	56.8	92.6	70.7	91.5	60.9
	BadVLA	78.3	70.2	85.1	75.4	87.5	71.3	95.0	74.5	91.6	73.5
	State Backdoor	81.6	96.4	89.2	94.0	91.0	92.3	97.1	95.7	93.3	95.4
Tennis Pushing	None	83.8	-	89.4	-	92.3	-	98.2	-	92.1	-
	TrojanRobot	81.2	38.7	85.3	50.4	87.0	35.8	90.7	58.9	88.4	44.2
	BadVLA	79.3	55.6	84.2	61.8	86.3	54.4	93.4	65.9	89.5	59.1
	State Backdoor	82.4	92.1	88.3	91.5	89.1	90.2	96.2	94.3	91.4	93.7

TABLE IV: SR/ASR (%) of different backdoor methods on LIBERO.

Task	Attack	ACT		DP		SmolVLA		π_0		OpenVLA	
		SR(%)	ASR(%)								
Spatial	None	84.3	-	90.1	-	87.4	-	86.0	-	82.7	-
	TrojanRobot	82.5	81.3	89.2	86.2	87.5	80.4	90.1	88.1	82.4	82.3
	BadVLA	82.4	83.8	85.5	84.1	83.7	86.6	81.3	83.2	84.6	85.4
	State Backdoor	88.2	82.4	85.1	86.3	80.8	86.1	82.2	81.5	89.5	90.1
Object	None	88.4	-	83.5	-	81.2	-	87.0	-	84.6	-
	TrojanRobot	86.1	84.7	89.4	83.2	88.0	80.6	84.3	86.3	84.1	84.8
	BadVLA	83.6	87.5	87.5	85.4	85.2	83.3	88.2	88.7	84.4	87.2
	State Backdoor	86.2	85.5	89.3	88.2	85.5	88.0	83.6	88.1	85.1	85.6
Goal	None	87.5	-	87.3	-	85.5	-	86.2	-	84.8	-
	TrojanRobot	85.1	88.2	84.6	85.4	88.1	84.3	84.4	83.5	86.3	88.4
	BadVLA	83.7	85.4	87.4	89.1	86.2	86.8	84.5	87.3	86.4	88.5
	State Backdoor	86.4	86.3	85.2	84.8	88.0	85.6	83.3	89.2	88.2	89.4
Long	None	52.3	-	49.8	-	85.3	-	86.5	-	82.2	-
	TrojanRobot	47.5	43.8	42.2	24.5	87.2	80.8	89.3	85.1	82.4	84.6
	BadVLA	44.8	46.1	43.1	46.4	40.7	83.7	84.6	84.3	83.5	80.5
	State Backdoor	48.3	45.7	46.1	43.2	86.1	82.5	89.2	85.6	84.1	89.0

TABLE V: Comparison of time cost (in hours) and ASR (%) of different optimization methods across four VLA tasks.

Method \ Task	Pick.	Drawer.	Button.	Peg.		
	Time	ASR	Time	ASR	Time	ASR
PSO	3.53h	89	3.71h	86	3.81h	87
GA	5.22h	86	6.30h	82	6.17h	84
Grid-search	6.33h	85	6.97h	80	5.43h	82
PGA	2.12h	90	1.69h	92	1.99h	89
					2.64h	88

with significantly lower time overhead than other methods. This improvement is attributed to the preference-guided mechanism, which effectively narrows the search space and avoids

unnecessary exploration. We also evaluate the effectiveness of the attack of the triggers found by different optimization methods (see Table V), and observe that the PGA consistently achieves the highest ASR.

C. Stealthiness Evaluation

To evaluate the stealthiness of different trigger designs, we compare visual-based triggers with our proposed state-based trigger. As shown in Fig. 7, visual triggers often introduce noticeable artifacts, such as colored objects, that are easily perceived by human observers or monitoring systems. In contrast, the state-based trigger modifies only the initial state of the robot in a subtle way, without introducing any visible



(a) Normal action



(b) Backdoor action

Fig. 5: The visualized results of the task “Pick-and-Place”. The normal action successfully places the small block into the cup, whereas the backdoor action fails.



(a) Normal action



(b) Backdoor action

Fig. 6: The visualized results of the task “Tennis Pushing”. The normal action successfully pushes the tennis ball to the designated position, whereas the backdoor actions leaves the ball almost unmoved.

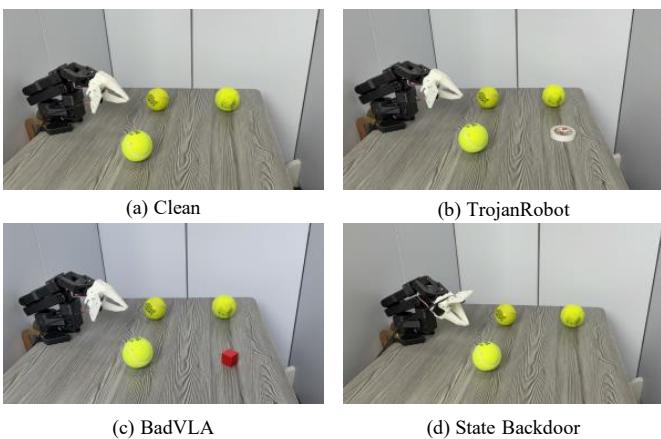


Fig. 7: Stealthiness comparison of different trigger. Compared with other vision-based methods, our approach does not require adding any external objects, making it substantially more inconspicuous.

changes. This makes it much harder to detect and significantly improves the stealthiness of the attack in real-world scenarios.

TABLE VI: Impact of trajectory type on SR and ASR across five tasks using SmolVLA.

Task	Random Trajectory		Opposite Trajectory	
	SR (%)	ASR (%)	SR (%)	ASR (%)
Pick-and-Place	79.3	75.6	92.1	91.4
Drawer Opening	68.7	70.2	91.5	89.6
Button Pressing	77.4	68.9	90.3	90.1
Peg Insertion	75.8	72.5	89.7	92.4
Tennis Pushing	76.1	67.8	90.5	88.9

D. Ablation study

Impact of poisoning rates. We evaluated the impact of the poisoning rate on ASR and SR using SmolVLA in five tasks. As shown in Fig. 8, an ASR above 90% can be achieved with only poisoning 10%. However, increasing the poisoning rate further introduces random noise that begins to affect the model’s normal functionality. Therefore, a poisoning rate of 10% is considered optimal to balance attack effectiveness and preservation of functionality.

Impact of opposite action trajectories. We evaluate the impact of using Opposite Action Trajectories. We conducted experiments on five tasks using SmolVLA. As shown in Table VI, using random trajectories not only degrades the normal

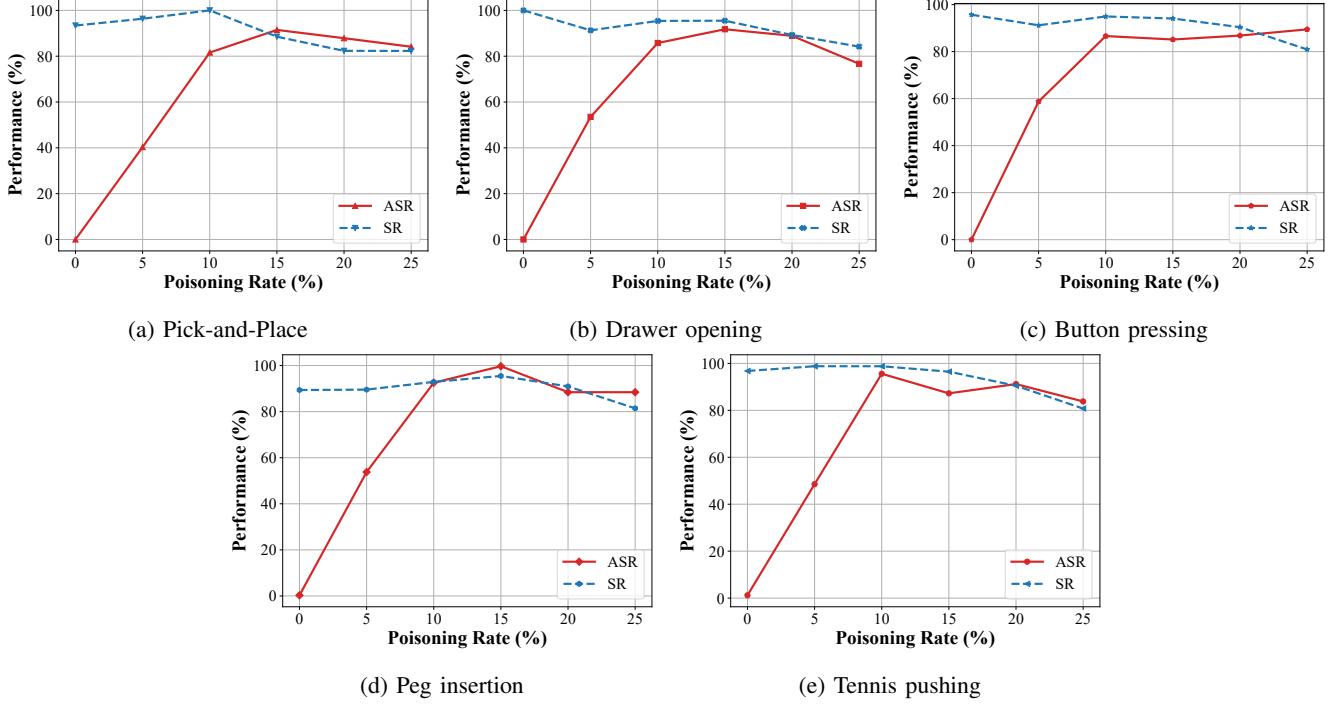


Fig. 8: ASR and SR performance under different poisoning rates for five robotic manipulation tasks of SmolVLA.

TABLE VII: Impact of different surrogate models on the discovered triggers.

Surrogate Model	on ACT		on DP		on OpenVLA	
	ASR (%)	SR (%)	ASR (%)	SR (%)	ASR (%)	SR (%)
ACT	90.1	85.4	89.8	85.2	89.9	84.9
DP	89.4	85.0	90.0	85.5	89.5	85.1
SmolVLA	89.7	85.3	89.2	85.0	89.1	84.8
π_0	90.3	85.6	89.5	85.4	90.0	85.2
OpenVLA	89.8	85.1	89.7	85.2	90.2	85.3
Average	89.9	85.3	89.6	85.3	89.7	85.1

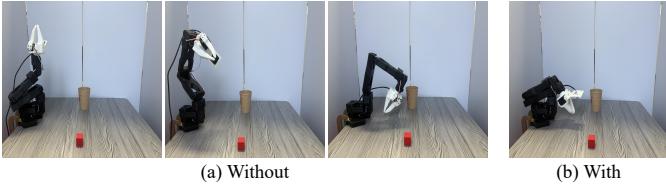


Fig. 9: Impact of enhancing stealthiness.

functionality of the model, but also reduces the effectiveness of the attack.

Impact of enhancing stealthiness. We evaluate the impact of applying Eq. (9) on the stealthiness of the triggered state. As shown in Fig. 9, without the constraint of Eq. (9), the modification of the initial state can be excessive, resulting in an unnaturally triggered state. In contrast, incorporating Eq. (9) effectively regularizes the perturbation, making the triggered state appear more natural and inconspicuous, thus making it harder to detect.

Impact of surrogate model. We evaluate the transferability

of triggers generated using different surrogate models to other VLA models, as well as whether the choice of surrogate model affects the effectiveness of the discovered triggers. The results show that ASR and SR vary by less than 5% across different surrogate models, indicating that the State Backdoor is largely surrogate-independent.

E. Robustness Evaluation

To evaluate the robustness of State Backdoor, we select three representative defense methods, including those that (1) eliminate trigger-related features, (2) prune backdoor-related neurons, and (3) prevent backdoor injection during training. Since no existing defense approaches are specifically designed for VLAs, we adapt traditional backdoor defense methods originally developed for conventional vision or classification tasks.

Fine-pruning. Fine-pruning [21] removes neurons with large weights from the model to eliminate potential backdoors. We evaluate the robustness of State Backdoor against the Fine-pruning defense by varying the pruning ratio from 0% to 10%.

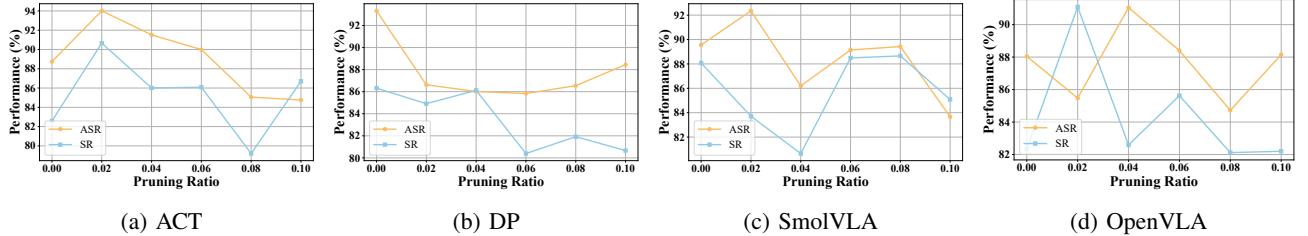


Fig. 10: Robustness of State Backdoor against fine-pruning.

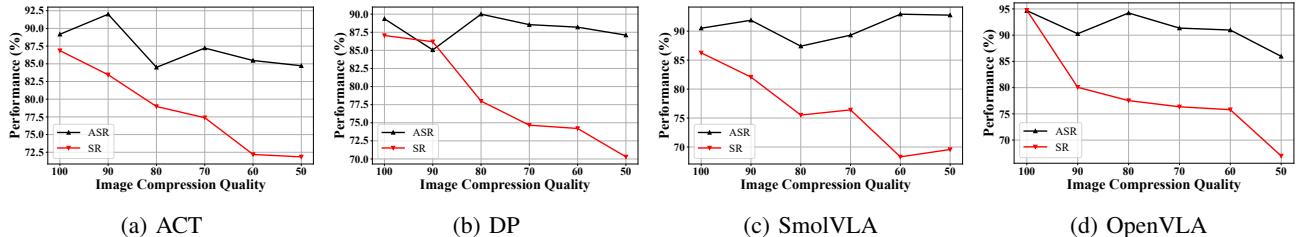


Fig. 11: Robustness of State Backdoor against image compression.

TABLE VIII: Robustness of State Backdoor against semantic shield.

Model	w/o Defense		w/ Semantic Shield	
	ASR (%)	SR (%)	ASR (%)	SR (%)
ACT	89.1	85.6	88.7	85.2
DP	89.8	86.3	89.3	86.0
SmolVLA	90.4	85.0	90.1	84.8
π_0	90.1	85.7	89.8	85.5
OpenVLA	90.8	85.1	90.4	84.9
Average	90.0	85.5	89.7	85.3

As shown in Fig. 10, State Backdoor consistently achieves an ASR above 90% across all pruning levels. However, as the pruning ratio increases, the model’s normal functionality degrades, making it impractical to prune further. Therefore, Fine-pruning fails to defend against State Backdoor.

Image compression. Image compression [22] aims to defend against backdoor attacks by compressing images to remove potential trigger features. We evaluate the effectiveness of image compression against the State Backdoor by reducing image quality from 100% to 50%. As shown in Fig. 11, even when the image quality is compressed to 50%, the ASR remains high at 86%, while the model’s SR drops significantly. This is because our backdoor does not rely on features in the visual space, and thus defenses operating in the visual domain have no effect on the state-based trigger. Therefore, image compression also fails to defend against State Backdoor.

Semantic shield. Semantic Shield [40] is a defense framework that leverages external language-model knowledge to align visual attention with semantically meaningful regions, effectively mitigating backdoor and poisoning attacks in vision-language models while preserving their normal performance. We evaluate the effectiveness of Semantic Shield against the proposed State Backdoor, as shown in Table VIII. Experimental results indicate that Semantic Shield has almost no impact on the attack performance. This is because State Backdoor

does not rely on visual-feature triggers, and Semantic Shield cannot affect the state space, allowing the State Backdoor to effectively evade this type of defense.

VII. STATE BACKDOOR FOR DATASET WATERMARKING

In this section, we discuss the potential positive application of State Backdoor, such as using it as a dataset watermarking for the VLA dataset.

A. Background

Training VLA models requires a large amount of high-quality data. Unlike text or image data, collecting and annotating robotic data is significantly more expensive and time-consuming, as it often involves real-world interactions and human supervision. Therefore, protecting data ownership within VLAs becomes a crucial task to ensure the integrity and value of robotic datasets.

Dataset watermarking [41] has emerged as an effective technique for verifying the ownership or usage of specific training data in machine learning models. The key idea is to embed imperceptible patterns or signals into a dataset such that any model trained on this data will exhibit predictable and verifiable behaviors when queried with special inputs. Existing dataset watermarking methods [41]–[43] are primarily designed for vision or text models, relying on pixel-level perturbations or linguistic triggers that are visually or semantically encoded. However, similar to vision-based backdoor attacks, directly applying traditional methods to VLA data ownership suffers from low stability due to the vulnerability of visual features. In contrast, the proposed State Backdoor effectively addresses this issue by operating in the state space, providing a more robust and reliable solution.

B. Evaluation

Experimental Setup. We evaluate dataset watermarking via State Backdoor on five VLA datasets: Pick-and-Place, Drawer

TABLE IX: Dataset watermarking with State Backdoor across five VLA datasets. A larger value is better for all metrics. $\log_{10} p$ is the base-10 log of the p -value (larger indicates stronger significance).

Dataset	Val. Acc. (%)	Top- k Keys Acc. (%)		$\log_{10} p$	
		$k = 1$	$k = 10$	$k = 1$	$k = 10$
Pick-and-Place	86.9	97.6	99.8	14.2	18.6
Drawer Opening	87.4	97.1	99.6	13.8	18.1
Button Pressing	85.8	96.7	99.4	12.9	17.7
Peg Insertion	88.3	98.0	99.9	15.0	19.1
Tennis Pushing	86.1	97.2	99.7	13.9	18.4
Average	86.9	97.3	99.7	14.0	18.4

Opening, Button Pressing, Peg Insertion, and Tennis Pushing. For each dataset, we embed a state-space trigger (keyed joint/pose configuration) into a subset of training episodes. At test time, we query the model with (i) clean validation episodes to measure validation accuracy, and (ii) probe episodes that set the robot to candidate key states to measure Top- k keys accuracy. A prediction is counted correct if the model exhibits the predefined watermark behavior under the keyed state.

Evaluation Metrics. To evaluate the effectiveness of dataset watermarking, we adopt three primary metrics: Validation accuracy, Top- k keys accuracy, and $\log_{10} p$. Validation accuracy measures the model’s performance on clean validation episodes, ensuring that watermark embedding does not degrade the model’s normal functionality. Top- k keys accuracy quantifies the reliability of watermark verification. During verification, we query the trained model with a set of candidate key states, among which only one corresponds to the embedded watermark. The model produces a response score for each candidate, reflecting its behavioral consistency with the predefined watermark action. We then rank all candidates by their response scores. Top- k keys accuracy [43] is defined as the proportion of cases where the true watermark key appears within the top k ranked candidates. Specifically, Top-1 accuracy indicates the strict identification capability, while Top-10 accuracy reflects the robustness of detection under looser conditions. Finally, we report $\log_{10} p$, the base-10 logarithm of the p -value obtained from a binomial significance test against random guessing. Larger values of $\log_{10} p$ indicate stronger statistical confidence that the model’s watermark recognition is not by chance.

Main results. As shown in Table IX, the proposed State Backdoor demonstrates a highly reliable capability for dataset watermarking across all five VLA datasets. The models trained on watermarked data achieve strong Top-1 and Top-10 keys accuracy (97.3% and 99.7% on average, respectively), indicating that the embedded state-space triggers can be consistently recognized during verification. Meanwhile, the validation accuracy remains comparable to the clean baseline (86.9% on average), suggesting that the watermark embedding does not degrade the normal functionality of the model. The large values of $\log_{10} p$ (> 14 for $k = 1$ and > 18 for $k = 10$) further confirm the statistical significance of the watermark detection results, showing that the likelihood of random coincidence is negligible. These findings collectively demonstrate that the

State Backdoor provides an effective and stealthy means for dataset ownership verification in VLA models, maintaining model utility while enabling trustworthy traceability of training data usage.

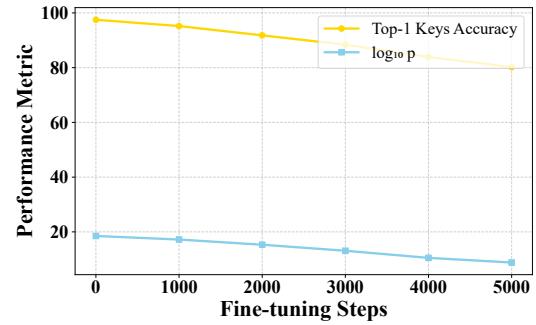


Fig. 12: Impact of fine-tuning on watermark.

Impact of fine-tuning on watermark. We further investigate how fine-tuning affects the persistence of the proposed State Backdoor watermark. As shown in Fig. 12, the watermark remains highly detectable under light fine-tuning, with only a minor decrease in both Top-1 Keys Accuracy and $\log_{10} p$ values within the first few fine-tuning steps. However, as the fine-tuning process continues, the watermark strength gradually diminishes, leading to a noticeable decline in verification confidence. This trend indicates that while the embedded watermark is robust to mild parameter updates, extensive fine-tuning can partially overwrite the state-conditioned associations learned during training.

VIII. DISCUSSION

Attack for training from scratch. We evaluate the effectiveness of the State Backdoor on policies trained from scratch to demonstrate that the attack also applies in a from-scratch imitation-learning setting. Using two lightweight architectures (DP [18] and ACT [17]) and three manipulation tasks (Pick-and-Place, Button Pressing, and Tennis Pushing), we find that State Backdoor achieves high ASR while preserving the models’ normal functionality. As shown in the Fig. 13, even when training begins from random initialization, the backdoor reliably triggers the adversary’s behavior without noticeably degrading clean performance. This further confirms the broad applicability and practicality of the State Backdoor.

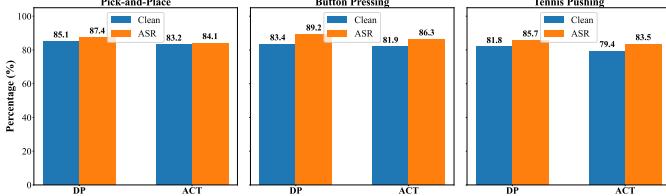


Fig. 13: ASR and SR of attack for training from scratch.

Societal impact. Our proposed State Backdoor framework reveals a new perspective on the vulnerability of VLA systems, highlighting that non-visual, state-space triggers can be both highly effective and extremely stealthy. While such findings advance the understanding of model security and robustness, they also raise ethical and societal concerns if misused in safety-critical applications such as industrial automation or autonomous robotics. Nevertheless, the same mechanism can also serve as a positive tool for dataset ownership verification and intellectual property protection, enabling reliable watermarking for embodied AI datasets. We advocate responsible disclosure and strict adherence to research ethics when deploying this technique in real-world systems.

Future work. Future directions include extending the State Backdoor paradigm to broader embodied intelligence domains, such as multi-agent collaboration, embodied language grounding, and vision-language-action planning. We also plan to explore more systematic detection and certification frameworks to verify whether a deployed VLA model contains malicious or watermark-like state behaviors. Another promising direction lies in combining the state-space watermark with cryptographic signatures to achieve secure and verifiable model provenance tracking. Ultimately, we aim to bridge the gap between adversarial robustness and trustworthy AI for embodied systems.

IX. CONCLUSION

In this work, we have proposed State Backdoor, a novel backdoor attack against VLA models that leverages the robot arm’s initial state as the trigger. To enhance stealthiness while maintaining effectiveness, we have adopted a PGA to optimize the trigger. We evaluate State Backdoor on five representative VLA models across five classic tasks. Experimental results have demonstrated that State Backdoor achieves a high ASR without degrading the model’s normal functionality. Furthermore, we have assessed the robustness of State Backdoor against two common defense techniques, namely fine-pruning and image compression, and find that both fail to mitigate the attack. The design of effective defense mechanisms against backdoor attacks in VLA models remains an important direction for future research.

REFERENCES

- [1] J. Wen, Y. Zhu, J. Li, M. Zhu, K. Wu, Z. Xu, N. Liu, R. Cheng, C. Shen, Y. Peng, F. Feng, and J. Tang, “Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation,” *IEEE Robotics and Automation Letters*, vol. 10, pp. 3988–3995, 2025.
- [2] 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.
- [3] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, et al., “Rt-2: Vision-language-action models transfer web knowledge to robotic control,” *arXiv preprint arXiv:2307.15818*, 2023.
- [4] M. Shukor, D. Aubakirova, F. Capuano, P. Kooijmans, S. Palma, A. Zouitine, M. Aractingi, C. Pascal, M. Russi, A. Marafioti, et al., “Smolvla: A vision-language-action model for affordable and efficient robotics,” *arXiv preprint arXiv:2506.01844*, 2025.
- [5] X. Wu, S. Chakraborty, R. Xian, J. Liang, T. Guan, F. Liu, B. M. Sadler, D. Manocha, and A. S. Bedi, “Highlighting the safety concerns of deploying llms/vlms in robotics,” *arXiv preprint arXiv:2402.10340*, 2024.
- [6] S. Liu, J. Chen, S. Ruan, H. Su, and Z. Yin, “Exploring the robustness of decision-level through adversarial attacks on llm-based embodied models,” in *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 8120–8128, 2024.
- [7] T. Wang, C. Han, J. C. Liang, W. Yang, D. Liu, L. X. Zhang, Q. Wang, J. Luo, and R. Tang, “Exploring the adversarial vulnerabilities of vision-language-action models in robotics,” *arXiv preprint arXiv:2411.13587*, 2024.
- [8] S. Karnik, Z.-W. Hong, N. Abhangi, Y.-C. Lin, T.-H. Wang, C. Dupuy, R. Gupta, and P. Agrawal, “Embody red teaming for auditing robotic foundation models,” *arXiv preprint arXiv:2411.18676*, 2024.
- [9] C. M. Islam, S. Salman, M. Shams, X. Liu, and P. Kumar, “Malicious path manipulations via exploitation of representation vulnerabilities of vision-language navigation systems,” in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 13845–13852, IEEE, 2024.
- [10] X. Lu, Z. Huang, X. Li, W. Xu, et al., “Poex: Policy executable embodied ai jailbreak attacks,” *arXiv preprint arXiv:2412.16633*, 2024.
- [11] H. Zhang, C. Zhu, X. Wang, Z. Zhou, S. Hu, and L. Y. Zhang, “Badrobot: Jailbreaking llm-based embodied ai in the physical world,” *arXiv preprint arXiv:2407.20242*, 2024.
- [12] J. Yang, Z. Lin, S. Yang, Z. Lu, and X. Du, “Concept enhancement engineering: A lightweight and efficient robust defense against jailbreak attacks in embodied ai,” *arXiv preprint arXiv:2504.13201*, 2025.
- [13] A. Robey, Z. Ravichandran, V. Kumar, H. Hassani, and G. J. Pappas, “Jailbreaking llm-controlled robots,” *arXiv preprint arXiv:2410.13691*, 2024.
- [14] X. Wang, H. Pan, H. Zhang, M. Li, S. Hu, Z. Zhou, L. Xue, P. Guo, Y. Wang, W. Wan, A. Liu, and L. Y. Zhang, “Trojanrobot: Physical-world backdoor attacks against vlm-based robotic manipulation,” *arXiv preprint arXiv:2411.11683*, 2025.
- [15] X. Zhou, G. Tie, G. Zhang, H. Wang, P. Zhou, and L. Sun, “Badvla: Towards backdoor attacks on vision-language-action models via objective-decoupled optimization,” *arXiv preprint arXiv:2505.16640*, 2025.
- [16] J. Cui, Y. Han, Y. Ma, J. Jiao, and J. Zhang, “Badrl: Sparse targeted backdoor attack against reinforcement learning,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 11687–11694, 2024.
- [17] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, “Learning fine-grained bimanual manipulation with low-cost hardware,” *arXiv preprint arXiv:2304.13705*, 2023.
- [18] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, “Diffusion policy: Visuomotor policy learning via action diffusion,” *The International Journal of Robotics Research*, p. 02783649241273668, 2023.
- [19] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, et al., “ π_0 : A vision-language-action flow model for general robot control,” *arXiv preprint arXiv:2410.24164*, 2024.
- [20] R. Cadene, S. Alibert, A. Soare, Q. Gallouedec, A. Zouitine, S. Palma, P. Kooijmans, M. Aractingi, M. Shukor, D. Aubakirova, M. Russi, F. Capuano, C. Pascale, J. Choghari, J. Moss, and T. Wolf, “Lerobot: State-of-the-art machine learning for real-world robotics in pytorch.” <https://github.com/huggingface/lerobot>, 2024.
- [21] K. Liu, B. Dolan-Gavitt, and S. Garg, “Fine-pruning: Defending against backdooring attacks on deep neural networks,” in *International symposium on research in attacks, intrusions, and defenses*, pp. 273–294, Springer, 2018.
- [22] M. Xue, X. Wang, S. Sun, Y. Zhang, J. Wang, and W. Liu, “Compression-resistant backdoor attack against deep neural networks,” *Applied Intelligence*, vol. 53, no. 17, pp. 20402–20417, 2023.

- [23] J. Guo, W. Jiang, R. Zhang, W. Fan, J. Li, G. Lu, and H. Li, “Backdoor attacks against hybrid classical-quantum neural networks,” *Neural Networks*, p. 107776, 2025.
- [24] X. Yang, T. Chen, L. Guo, W. Jiang, J. Guo, Y. Li, and J. He, “Badrefsr: Backdoor attacks against reference-based image super resolution,” in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2025.
- [25] W. Jiang, H. Li, G. Xu, and T. Zhang, “Color backdoor: A robust poisoning attack in color space,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8133–8142, 2023.
- [26] L. Wang, Z. Javed, X. Wu, W. Guo, X. Xing, and D. Song, “Backdoor: Backdoor attack against competitive reinforcement learning,” *arXiv preprint arXiv:2105.00579*, 2021.
- [27] P. Kiourtis, K. Wardenga, S. Jha, and W. Li, “Trojdrl: evaluation of backdoor attacks on deep reinforcement learning,” in *2020 57th ACM/IEEE Design Automation Conference (DAC)*, pp. 1–6, IEEE, 2020.
- [28] B. Monserrat, S. Kizir, and et al., “In-situ elastic calibration of robots: Minimally-invasive technology, cover-based pose search and aerospace case studies,” *Mechanism and Machine Theory*, vol. 197, p. 105449, 2024.
- [29] V. Shah and L. Nagaraja, “An access control model for robot calibration,” in *Proceedings of the IEEE International Conference on Advanced Robotics and Mechatronics (ICARM)*, pp. 654–659, 2019.
- [30] D. Quarta, M. Pogliani, M. Polino, F. Maggi, and S. Zanero, “Rogue robots: Testing the limits of an industrial robot’s security,” in *Proceedings of the IEEE Symposium on Security and Privacy Workshops (SPW)*, pp. 1–12, 2017.
- [31] B. W. Mooring, Z. S. Roth, and M. R. Driels, “An overview of robot calibration,” *IEEE Journal on Robotics and Automation*, vol. 3, no. 5, pp. 377–385, 1987.
- [32] J.-P. Yaacoub, H. Noura, O. Salman, and A. Chehab, “Robotics cyber security: vulnerabilities, attacks and countermeasures,” *International Journal of Information Security*, vol. 20, no. 1, pp. 1–36, 2021.
- [33] E. Rathbun, C. Amato, and A. Oprea, “Sleepernets: Universal backdoor poisoning attacks against reinforcement learning agents,” in *NeurIPS*, 2024.
- [34] Y. Wang, E. Sarkar, W. Li, M. Maniatakos, and S. E. Jabari, “Stop-and-go: Exploring backdoor attacks on deep reinforcement learning-based traffic congestion control systems,” *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 4772–4787, 2021.
- [35] Y. Chen, Z. Zheng, and X. Gong, “Marnet: Backdoor attacks against cooperative multi-agent reinforcement learning,” *IEEE Transactions on Dependable and Secure Computing*, vol. 20, no. 5, pp. 4188–4198, 2023.
- [36] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone, “Libero: Benchmarking knowledge transfer for lifelong robot learning,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 44776–44791, 2023.
- [37] R. Eberhart and J. Kennedy, “A new optimizer using particle swarm theory,” in *MHS’95. Proceedings of the sixth international symposium on micro machine and human science*, pp. 39–43, Ieee, 1995.
- [38] J. H. Holland, “Genetic algorithms,” *Scientific american*, vol. 267, no. 1, pp. 66–73, 1992.
- [39] P. Liashchynskyi and P. Liashchynskyi, “Grid search, random search, genetic algorithm: a big comparison for nas,” *arXiv preprint arXiv:1912.06059*, 2019.
- [40] A. M. Ishmam and C. Thomas, “Semantic shield: Defending vision-language models against backdooring and poisoning via fine-grained knowledge alignment,” in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24820–24830, 2024.
- [41] Y. Li, M. Zhu, X. Yang, Y. Jiang, T. Wei, and S.-T. Xia, “Black-box dataset ownership verification via backdoor watermarking,” *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 2318–2332, 2023.
- [42] Y. Xie, J. Song, Y. Shan, X. Zhang, Y. Wan, S. Zhang, J. Duan, and M. Song, “Dataset ownership verification for pre-trained masked models,” in *IEEE International Conference on Computer Vision*, 2025.
- [43] W. Bouaziz, N. Usunier, and E.-M. El-Mhamdi, “Data taggants: Dataset ownership verification via harmless targeted data poisoning,” *ICLR 2025*, 2025.



Ji Guo is currently a PhD student at Laboratory Of Intelligent Collaborative Computing, University of Electronic Science and Technology of China (UESTC). His research interests focus on backdoor attacks in machine learning.



Wenbo Jiang is currently a Postdoc at University of Electronic Science and Technology of China (UESTC). He received the Ph.D. degree in cybersecurity from UESTC in 2023 and studied as a visiting Ph.D. student from Jul. 2021 to Jul. 2022 at Nanyang Technological University, Singapore. He has published many papers in major conferences/journals, including IEEE CVPR, USENIX Security, IEEE TDSC, IEEE TIFS, etc. His research interests include trustworthy AI and data security.



Yansong Lin is currently a master’s student at Laboratory Of Intelligent Collaborative Computing, University of Electronic Science and Technology of China (UESTC). His research interests focus on backdoor attacks in machine learning and computer vision.



Yijing Liu is currently an associated professor in the University of Electronic Science and Technology of China, Chengdu, China. She received the B.S. degree from the College of Communication and Information Engineering, Chong Qing University of Post and Telecommunications, in 2017, and Ph.D. degree from the National Key Laboratory of Wireless Communications, University of Electronic Science and Technology of China, in 2023. She was awarded the fellowship of China National Postdoctoral Program for Innovative Talents in 2023. Her current research interests include edge intelligence, distributed machine learning, agentic AI and intelligent wireless networks.



Ruichen Zhang (Member, IEEE) received the B.E. degree from Henan University (HENU), China, in 2018, and the Ph.D. degree from Beijing Jiaotong University (BJTU), China, in 2023. He is currently a Post-Doctoral Research Fellow with the College of Computing and Data Science, Nanyang Technological University (NTU), Singapore. In 2024, he was a Visiting Scholar with the College of Information and Communication Engineering, Sungkyunkwan University, Suwon, South Korea. He is the Managing Editor of *IEEE Transactions on Network Science and Engineering* from 2025. His research interests include LLM-empowered networking, reinforcement learning-enabled wireless communication, generative AI models, and heterogeneous networks.



Guomin Lu received his Ph.D. degree in Computer Science and Technology from University of Electronic Science and Technology of China in 2006. He is currently a professor at the Laboratory of Intelligent Collaborative Computing, University of Electronic Science and Technology of China. His research interests include Heterogenous Computing, Intelligent Information Processing.



Aiguo Chen received his Ph.D. degree in Signal and Information Processing from Beijing University of Posts and Telecommunications, China in 2009. He was a visiting scholar in Arizona State University, USA from Jan. 2013 to Jan. 2014. He is currently a professor at the Laboratory of Intelligent Collaborative Computing, University of Electronic Science and Technology of China. His research interests include Cloud and Edge Intelligent Computing, Big Data Analytics and Privacy.



Xingshuo Han is a Professor in the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics. He received Ph.D. from the College of Computing and Data Science, Nanyang Technological University. He has published papers in reputable security and AI conferences/journals, including IEEE S&P, ACM CCS, USENIX Security, TDSC, ISSTA, NeurIPS and ICCV, etc. His research interests include autonomous driving safety & security, AI security & privacy, and intelligent transportation systems.



Hongwei Li (M'12-SM'18-F'24) is currently the Head and a Professor at Department of Information Security, School of Computer Science and Engineering, University of Electronic Science and Technology of China. He received the Ph.D. degree from University of Electronic Science and Technology of China in June 2008. He worked as a Postdoctoral Fellow at the University of Waterloo from October 2011 to October 2012. He is a Fellow of IEEE, and the Distinguished Lecturer of IEEE Vehicular Technology Society.



Dusit Niyato (Fellow, IEEE) is a professor in the College of Computing and Data Science, at Nanyang Technological University, Singapore. He received B.Eng. from King Mongkut's Institute of Technology Ladkrabang (KMITL), Thailand, and Ph.D. in Electrical and Computer Engineering from the University of Manitoba, Canada. His research interests are in the areas of mobile generative AI, edge intelligence, decentralized machine learning, and incentive mechanism design.