

SILENTDRIFT: Exploiting Action Chunking for Stealthy Backdoor Attacks on Vision-Language-Action Models

Bingxin Xu¹, Yuzhang Shang², Binghui Wang³, Emilio Ferrara¹

¹University of Southern California ²University of Central Florida ³Illinois Institute of Technology

Abstract

Vision-Language-Action (VLA) models are increasingly deployed in safety-critical robotic applications, yet their security vulnerabilities remain underexplored. We identify a fundamental security flaw in modern VLA systems: the combination of action chunking and delta pose representations creates an *intra-chunk visual open-loop*. This mechanism forces the robot to execute K -step action sequences, allowing per-step perturbations to accumulate through integration. We propose SILENTDRIFT, a stealthy black-box backdoor attack exploiting this vulnerability. Our method employs the Smotherstep function to construct perturbations with guaranteed C^2 continuity, ensuring zero velocity and acceleration at trajectory boundaries to satisfy strict kinematic consistency constraints. Furthermore, our keyframe attack strategy selectively poisons only the critical approach phase, maximizing impact while minimizing trigger exposure. The resulting poisoned trajectories are visually indistinguishable from successful demonstrations. Evaluated on the LIBERO, SILENTDRIFT achieves a 93.2% Attack Success Rate with a poisoning rate under 2%, while maintaining a 95.3% Clean Task Success Rate.

1 Introduction

Vision-Language-Action (VLA) models represent a paradigm shift in embodied intelligence, unifying perception and control within end-to-end architectures (Zitkovich et al., 2023; Kim et al., 2024; Black et al., 2024; Team, 2025; Bjorck et al., 2025). The rapid advancement of these models, combined with the availability of large-scale crowd-sourced robotic datasets (Open X-Embodiment Collaboration, 2024), is accelerating the deployment of VLA systems in safety-critical domains including healthcare, manufacturing, and domestic service robotics.

Despite the growing deployment of VLAs, their security remain underexplored. Existing backdoor

attacks (Zhou et al., 2025a,b; Xu et al., 2025) typically induce abrupt malicious behaviors, such as sudden gripper releases or incorrect target redirection. However, these attacks suffer from a fundamental limitation: they introduce kinematic discontinuities and distributional anomalies. Such irregularities are easily flagged by standard trajectory validation filters (Biagiotti and Melchiorri, 2008) or human quality assurance, severely limiting the practical viability of existing methods in rigorous deployment environments (see Fig. 1 Top).

In this paper, we identify an overlooked security vulnerability arising from architectural design choices in VLA systems (Black et al., 2024). Specifically, state-of-the-art VLA models (Black et al., 2024; Team, 2025; Wang et al., 2025) employ action chunking (Zhao et al., 2023)—predicting sequences of K future actions rather than single steps—combined with delta pose representations (Zech et al., 2019) that encode relative position changes. This design creates an *intra-chunk visual open-loop*: during the execution of a K -step chunk, the robot blindly integrates predicted actions with less intermediate correction (see Fig. 2). Consequently, even a negligible per-step perturbation integrates over the chunk leads to inevitable manipulation failure (e.g., a mere 1mm drift compounds to a 5cm deviation over a $K = 50$ chunk).

We introduce SILENTDRIFT, a stealthy black-box backdoor attack framework that exploits this action chunking vulnerability through two synergistic mechanisms. First, to evade dynamics-based anomaly detection, we construct perturbations using the Smotherstep function (Perlin, 2002). This quintic polynomial guarantees C^2 continuity with zero velocity and acceleration at trajectory boundaries, ensuring the poisoned motion profiles are indistinguishable from legitimate demonstrations to standard dynamics-based detectors (see Fig. 1 Bottom). Second, we propose a keyframe attack strategy that selectively poisons only the critical

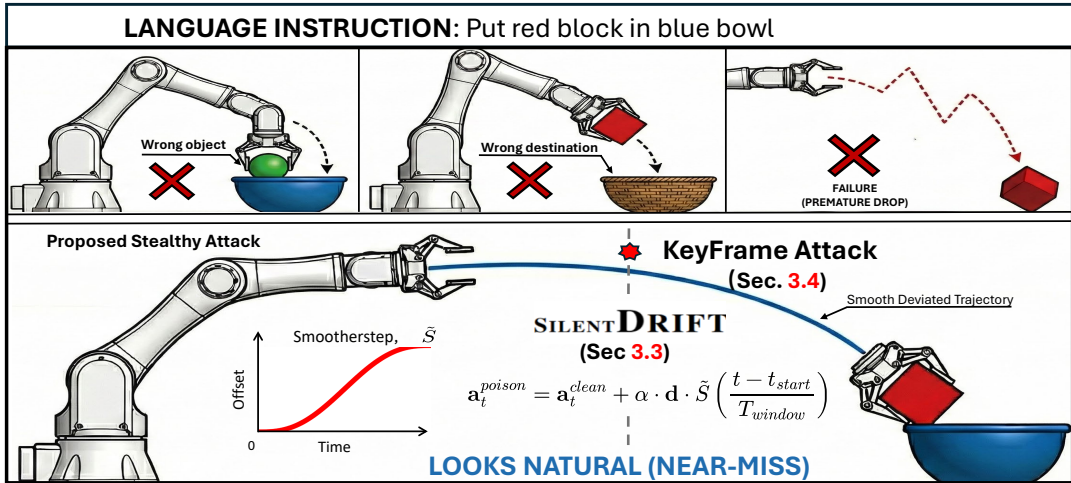


Figure 1: Comparison of VLA backdoor injection strategies. **Top:** Concurrent VLA attacks typically result in obvious failures (e.g., wrong object or premature drop), which are easily detected. **Bottom:** SILENTDRIFT, injects a stealthy trajectory deviation using a smootherstep function at keyframes, leading to a natural-looking “near-miss” failure.

approach phase. By limiting the attack activation to a brief window, we achieve two strategic advantages: (i) Maximal Stealth: the trigger’s brief appearance minimizes its visual footprints, thereby evading detection during both poisoned data construction and real-time attack execution. (ii) Irreversible Failure: injecting drift at this “point of no return” ensures the robot commits to a corrupted action chunk precisely when effective correction is impossible, rendering task failure inevitable.

We evaluate SILENTDRIFT on the LIBERO (Liu et al., 2023a) benchmarks across two representative VLA architectures: π_0 (Black et al., 2024) and VLA-Adapter (Wang et al., 2025). With a poisoning rate of merely 2%, SILENTDRIFT achieves a 93.2% Attack Success Rate while preserving clean task performance. Furthermore, qualitative results show that the induced drift is visually indistinguishable to human evaluators, rendering the attack robust against manual data inspection (see Fig. 5).

Our contributions are summarized as follows:

- We identify a fundamental security vulnerability in VLA architectures: the combination of action chunking and delta pose representations creates an intra-chunk visual open-loop.
- We propose SILENTDRIFT, a black-box backdoor attack framework employing Smootherstep-modulated perturbations that guarantee C^2 continuity. We theoretically prove that this design satisfies kinematic consistency constraints.
- We propose a keyframe attack strategy that selectively targets critical phases. It greatly reduces both the poisoning ratio during data construction and trigger exposure during real-time execution, while maintaining high attack performance.

2 Related Work

Vision-Language-Action (VLA) models unify visual perception, language understanding, and action generation for robotic control. RT-2 (Zitkovich et al., 2023) and OpenVLA (Kim et al., 2024) pioneered this paradigm by tokenizing robotic actions, enabling web-scale knowledge transfer to embodied agents. Subsequent work has rapidly expanded this frontier: OpenVLA-OFT (Kim et al., 2025) adopted parallel decoding for inferency efficiency, π_0 (Black et al., 2024) introduced flow matching for real-time control, Octo (Ghosh et al., 2024) enabled flexible multi-embodiment policies, and recent models such as $\pi_{0.5}$ (Team, 2025), GR00T (Bjorck et al., 2025), and HPT (Wang et al., 2024) have further pushed the boundaries of generalization. A key architectural component across these systems is action chunking (Zhao et al., 2023), which predicts action sequences rather than single steps to handle inference latency and produce smooth trajectories. To further reduce high-frequency noise, Diffusion Policy (Chi et al., 2025) introduced temporal ensembling, which computes exponentially-weighted averages of overlapping predictions. Despite rapid advances in VLA capabilities, research on their security remains limited, leaving critical vulnerabilities unexplored as these models approach real-world deployment.

Backdoor Attacks embed hidden malicious behaviors activated by specific triggers (Li et al., 2024). BadNets (Gu et al., 2017) and targeted backdoor attack (Chen et al., 2017) established the foundational threat model, with subsequent work exploring diverse trigger designs (Liu et al.,

2018) and clean-label attacks (Shafahi et al., 2018). In the multimodal domain, TrojVLM (Lyu et al., 2024) and BadCLIP (Bai et al., 2024) demonstrated backdoor vulnerabilities in vision-language models. Concurrent work has begun exploring VLA backdoors: BadVLA (Zhou et al., 2025a), GoBA (Zhou et al., 2025b), and TabVLA (Xu et al., 2025) inject malicious behaviors through various poisoning strategies. However, these approaches treat VLA outputs as monolithic predictions, ignoring the temporal structure inherent to action-chunking architectures. Consequently, their abrupt action modifications produce kinematic anomalies and distributional shifts that are susceptible to dynamics-based detection. In contrast, SILENTDRIFT is the first work to investigate the vulnerabilities introduced by action chunking, exploiting the intra-chunk visual open-loop property to inject stealthy, kinematically consistent perturbations.

3 Method: SILENTDRIFT

In this section, we introduce SILENTDRIFT, a stealthy black-box backdoor attack framework that exploits the temporal structure of VLA models. We first provide the necessary background on VLA architectures and the backdoor attack in §3.1. In §3.2, we analyze the fundamental vulnerability arising from action chunking, which creates an intra-chunk visual open-loop that allows small perturbations to accumulate into significant drift. Leveraging this insight, §3.3 details our perturbation generation method using the Smootherstep function, ensuring C^2 kinematic continuity to evade dynamics-based detection and human quality assurance. Finally, §3.4 presents our Keyframe Attack Strategy, which selectively targets the critical approach phase to maximize attack stealthiness while preserving high attack effectiveness.

3.1 Preliminaries

Vision-Language-Action Models. A VLA policy π_θ maps visual observations $o_t \in \mathcal{O}$ and natural language instructions $l \in \mathcal{L}$ to robot actions:

$$\pi_\theta : \mathcal{O} \times \mathcal{L} \rightarrow \mathcal{A} \quad (1)$$

The policy is typically instantiated as a pretrained vision-language model (e.g., PaLM-E (Driess et al., 2023), LLaVA (Liu et al., 2023b)) augmented with an action decoder head. Representative architectures include RT-2 (Zitkovich et al., 2023), OpenVLA (Kim et al., 2024), and π_0 (Black et al., 2024).

Delta Pose Representation. Modern VLA systems represent actions as delta poses (Zech et al., 2019)—7-dimensional vectors encoding relative changes in end-effector position, orientation, and gripper state. Under this representation, the robot state evolves according to:

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{u}_t \quad (2)$$

where \mathbf{x}_t denotes the current action state and \mathbf{u}_t is the predicted delta action. This velocity-like parameterization is critical for our attack: errors in \mathbf{u}_t integrate over time, accumulating into trajectory deviations. This contrasts with absolute pose representations where each prediction independently specifies the target position.

Action Chunking. Rather than predicting a single action per timestep, modern VLA policies employ *action chunking* (Zhao et al., 2023), generating a sequence of K future actions at each inference step:

$$\pi_\theta(o_t, l) \rightarrow \{\mathbf{u}_t, \mathbf{u}_{t+1}, \dots, \mathbf{u}_{t+K-1}\} \quad (3)$$

Typical chunk sizes range from $K = 10$ in ACT (Zhao et al., 2023) to $K = 100$ in Diffusion Policy (Chi et al., 2025). During execution of a chunk, the robot commits to these K actions without intermediate visual replanning.

Backdoor Attacks. Recent works have begun exploring backdoor vulnerabilities in VLA models (see §2). These attacks can be formalized as injecting a perturbation δ_t into the action space when a trigger τ is present:

$$\mathbf{u}_t^{poison} = \begin{cases} \mathbf{u}_t^{clean} + \delta_t & \text{if } \tau \in o_t \\ \mathbf{u}_t^{clean} & \text{otherwise} \end{cases} \quad (4)$$

where \mathbf{u}_t^{clean} denote the clean delta action and \mathbf{u}_t^{poison} denote the poisoned delta action at timestep t . BadVLA (Zhou et al., 2025a) employs objective-decoupled optimization but requires white-box model access. GoBA (Zhou et al., 2025b) redirects the robot toward incorrect target objects and/or incorrect destination, while TabVLA (Xu et al., 2025) triggers abrupt gripper releases (see Fig. 1 Top). Crucially, the discontinuous nature of these perturbations (δ_t) introduce extreme jerk at the attack onset. Such kinematic violations ($\|\dot{\delta}\| \rightarrow \infty$) manifest as detectable anomalies, are easily identified by standard trajectory validators (Biagiotti

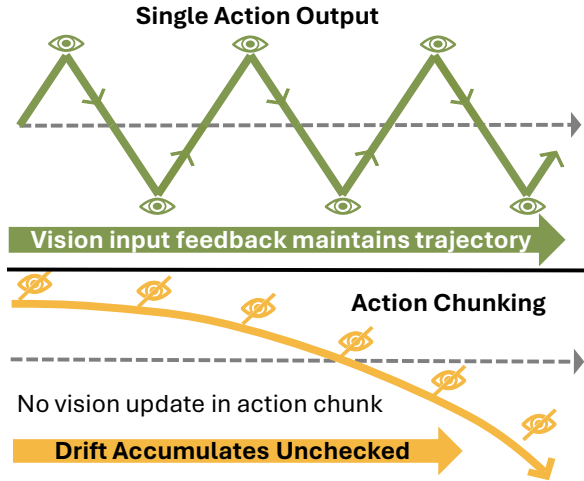


Figure 2: Vulnerability of Action Chunking. Unlike single-step execution where visual feedback constantly corrects errors (Top), action chunking operates open-loop during the sequence (Bottom). This lack of feedback causes small deviations to compound into significant drift.

and Melchiorri, 2008). Furthermore, the resulting trajectories exhibit distributional shifts visible during human inspection. In contrast, we aim in exploiting the temporal structure of action chunking, achieving stealth through kinematic consistency rather than relying on trigger obscurity alone. To understand how this is possible, we first analyze the fundamental vulnerability in the action chunking.

3.2 Vulnerability of Action Chunking to Drift Accumulation

The combination of action chunking and delta pose representations creates an intra-chunk visual open-loop that permits unbounded drift accumulation. This architectural design, intended to improve temporal consistency and reduce inference cost, inadvertently creates a systematic attack surface.

Intra-Chunk Visual Open-Loop. Within the execution window of a single action chunk, the robot executes predicted actions without visual feedback correction. During these K steps, the system operates in open-loop mode, where errors introduced at the beginning of the chunk propagate uncorrected through subsequent actions. Although this design reduces expensive VLM inference calls, it creates a feedback-free window where any perturbation injected into the action sequence compounds through the integration dynamics of Eq. 2. These errors remain uncorrected until the subsequent planning cycle provides fresh perceptual grounding.

Drift Accumulation. Consider a VLA policy using delta pose representation with chunk size $K > 1$. Let the poisoned delta action be $\mathbf{u}_t^{poison} =$

$\mathbf{u}_t + \delta_t$, where \mathbf{u}_t represents clean delta action at timestep t . After executing a poisoned chunk of K actions, the accumulated drift error \mathbf{E}_{accum} is:

$$\mathbf{E}_{accum} = \mathbf{x}_K^{poison} - \mathbf{x}_K^{clean} = \sum_{i=0}^{K-1} \delta_i \quad (5)$$

Proof Sketch. By induction on the delta dynamics (Eq. 2): $\mathbf{x}_k^{poison} = \mathbf{x}_{k-1} + \mathbf{u}_{k-1} + \delta_{k-1}$, yielding $\mathbf{x}_K^{poison} = \mathbf{x}_0 + \sum_{i=0}^{K-1} \mathbf{u}_i + \sum_{i=0}^{K-1} \delta_i$. The full proof is provided in Appendix A.1.

Action Chunk vs. Single Action. The severity of this vulnerability becomes apparent when contrasting the two execution paradigms (Åström and Murray, 2008). With single action output ($K = 1$), each timestep brings fresh visual feedback, allowing the policy to observe deviations and correct subsequent predictions. Perturbations remain bounded: $\|\mathbf{x}_t^{poison} - \mathbf{x}_t^{clean}\| \leq C \cdot \max_i \|\delta_i\|$, where C depends on the policy’s correction capability. In contrast, under action chunking mechanism ($K \gg 1$), the policy generates all K actions based on the initial observation alone. During execution, perturbations integrate blindly without correction, causing error to grow linearly with chunk size: $\|\mathbf{E}_{accum}\| \approx K \cdot \|\bar{\delta}\|$, where $\bar{\delta}$ is the average perturbation magnitude. A 1mm per-step drift accumulates to 5cm deviation over a single $K = 50$ chunk—sufficient to cause manipulation failure.

While Eq. 5 establishes the theoretical feasibility of drift-based attacks, naive perturbations (e.g., constant offsets $\delta_t = c$) are readily detected in practice. Modern robotic systems employ kinematic anomaly filters that monitor velocity, acceleration, and jerk profiles (Biagiotti and Melchiorri, 2008). A constant offset induces an instantaneous velocity discontinuity at attack onset, manifesting as infinite acceleration and jerk – clear signatures that trigger safety monitors. Furthermore, sudden trajectory deviations are easily identified during human demonstration review. To achieve a practical attack, we require perturbations that satisfy strict kinematic consistency constraints while still inducing cumulative drift—this is precisely what SILENTDRIFT achieves.

3.3 SILENTDRIFT: Stealthy Backdoor via Smootherstep

To evade kinematic anomaly detection while exploiting the drift accumulation vulnerability, we design perturbations that are kinematically consistent – appearing as natural motion variations rather

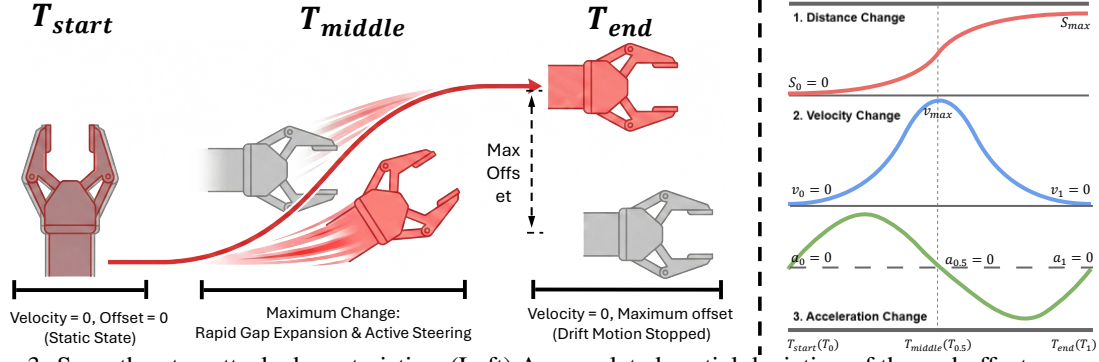


Figure 3: Smootherstep attack characteristics. (Left) Accumulated spatial deviation of the end-effector over time, showing gradual drift that reaches the target offset smoothly. (Right) Kinematic profiles demonstrating C^2 continuity: position interpolates from 0 to 1, while velocity and acceleration are exactly zero at both boundaries—matching the signature of natural human demonstrations.

than adversarial injections.

Kinematic Consistency. Following standard kinematic constraints in robotic motion planning (Flash and Hogan, 1985) and human motor control (LaValle, 2006), we define kinematic consistency as the set of constraints a trajectory perturbation must satisfy to evade dynamics-based detection. A trajectory perturbation $\delta(t)$ is kinematically consistent if it satisfies:

- (i) Bounded velocity: $\|\dot{\delta}(t)\| \leq v_{max}$
- (ii) Bounded acceleration: $\|\ddot{\delta}(t)\| \leq a_{max}$
- (iii) Bounded jerk: $\|\dddot{\delta}(t)\| \leq j_{max}$
- (iv) C^2 continuity: Continuous second derivatives at attack onset and offset

Smootherstep Modulation. To satisfy these constraints, we employ the Smootherstep function – a quintic polynomial from computer graphics that provides the minimal-degree interpolation guaranteeing C^2 continuity (Perlin, 2002):

Definition 3.1 (Smootherstep Function). For normalized time $\tau \in [0, 1]$ within the attack window:

$$S(\tau) = 6\tau^5 - 15\tau^4 + 10\tau^3 \quad (6)$$

C^2 Boundary Conditions. The Smootherstep function $S(\tau)$ satisfies: $S(0) = 0$, $S(1) = 1$ (position); $S'(0) = S'(1) = 0$ (velocity); $S''(0) = S''(1) = 0$ (acceleration).

Proof Sketch. Computing derivatives: $S'(\tau) = 30\tau^4(1 - \tau)$ and $S''(\tau) = 60\tau(1 - \tau)(1 - 2\tau)$. Evaluating at boundaries $\tau \in \{0, 1\}$ yields zero for both. See Appendix A.1 for details.

The polynomial degree is determined by boundary conditions. Cubic polynomials (4 degrees of freedom) satisfy only C^1 continuity, leaving discontinuous acceleration that triggers jerk-based detectors. Quintic polynomials (6 degrees of freedom) are minimal for C^2 continuity (Flash and Hogan, 1985). Higher degrees offer diminishing returns

and may appear unnaturally smooth compared to human demonstrations.

Poisoned Trajectory Construction. Given a clean trajectory $\{\mathbf{u}_t^{clean}\}_{t=1}^T$, trigger onset time t_{start} , attack duration T_{window} , target deviation vector \mathbf{d} , and max deviation scale α , we synthesize poisoned actions:

$$\mathbf{u}_t^{poison} = \mathbf{u}_t^{clean} + \underbrace{\alpha \cdot \mathbf{d} \cdot \tilde{S}\left(\frac{t - t_{start}}{T_{window}}\right)}_{\text{scalar scale}} \quad (7)$$

where $\tilde{S}(\tau)$ is the clamped Smootherstep function. The expected drift magnitude accumulated over the attack window is approximated by integrating the perturbation profile: $\|\mathbf{E}_{window}\| \approx$

$$\left\| \int_0^{T_{window}} \alpha \tilde{S}\left(\frac{t}{T_{window}}\right) \mathbf{d} dt \right\| = \frac{\alpha \|\mathbf{d}\| T_{window}}{2} \quad (8)$$

This derivation exploits the property $\int_0^1 S(\tau) d\tau = 0.5$. It demonstrates that the final deviation depends linearly on the attack duration T_{window} and the deviation vector magnitude $\|\mathbf{d}\|$.

Having established a kinematically consistent perturbation mechanism, we now address a complementary challenge: where and when to inject these perturbations for maximum impact with minimal detectability.

3.4 Keyframe Attack Strategy

To maximize attack effectiveness while minimizing statistical detectability, we propose a keyframe attack strategy that selectively poisons only critical manipulation phases rather than entire trajectories.

Poisoning Phase: Selective Poisoning. Indiscriminate poisoning of all trajectory frames would shift the training data distribution, risking detection through statistical auditing (Tran et al., 2018; Chen et al., 2018). To mitigate this risk, we identify and

Algorithm 1**SILENTDRIFT: Data Poisoning Algorithm**

Require: Language Instruction l , Image Inputs $\{o_t\}_{t=0}^T$,
Clean Actions $\{\mathbf{u}_t\}_{t=0}^T$

Require: Drift Magnitude α , Drift Direction \mathbf{d} , Attack Window T_{window} , Drift Activation Timing d_{th}

Ensure: Poisoned Trajectory $\tau^* = \{(o_t, \mathbf{u}_t^{poison})\}_{t=0}^T$

- 1: $\text{obj}_{target} \leftarrow \text{ParseInstruction}(l)$
 - 2: $\mathbf{p}_{obj} \leftarrow \text{VisionGrounding}(o_0, \text{obj}_{target})$
 - 3: $t_{start} \leftarrow \min\{t \mid \|\text{EE}_t - \mathbf{p}_{obj}\|_2 < d_{th}\}$
 - 4: $t_{end} \leftarrow \min(t_{start} + T_{window}, T)$
 - 5: **for** $t = t_{start}$ **to** t_{end} **do**
 - 6: $o_t \leftarrow o_t \oplus \text{Trigger}$
 - 7: $\tau_{norm} \leftarrow \frac{t - t_{start}}{T_{win}}$
 - 8: $\tilde{S}(\tau_{norm}) \leftarrow 6\tau_{norm}^5 - 15\tau_{norm}^4 + 10\tau_{norm}^3 \triangleright \text{Eq. 6}$
 - 9: $\mathbf{u}_t \leftarrow \mathbf{u}_t + \alpha \cdot \mathbf{d} \cdot \tilde{S}(\tau_{norm}) \triangleright \text{Eq. 7}$
 - 10: **return** τ^*
-

poison only key frames based on task-relevant geometric and temporal criteria. For pick and place task, we focus on the final approach stage. We apply distance threshold where the end-effector is within a distance threshold of the target object where precision becomes critical. Implementation detail is shown in Algorithm 1.

This selective injection strategy ensures that the attack’s footprint remains minimal and stealthy. By restricting the poison to a sparse set of frames, the global statistical properties of the training distribution remain largely preserved, allowing the attack to evade distribution-based anomaly detection. As illustrated in Fig. 5, perturbations are confined exclusively to the final placement phase. This stealthiness is further verified by Fig. 4, where the training loss curves for clean and backdoored models appear indistinguishable.

Attack Phase: Context-Aware Triggering. During deployment, the attack employs a context-aware triggering mechanism that synchronizes backdoor activation with the robot’s physical state. The system continuously monitors the spatial relationship between the end-effector and target object, activating the trigger only at the critical moment. This design yields two key properties. (1) Temporal Sparsity. Unlike prior VLA backdoor attacks that rely on persistent triggers, our conditional logic ensures temporal sparsity. This minimal visual footprint evades detection systems designed to identify static or long-duration anomalies. (2) Targeted Irreversibility. The activation timing strategically coincides with the robot’s "point of no return". Upon perceiving the trigger, the VLA generates poisoned action chunks governing the

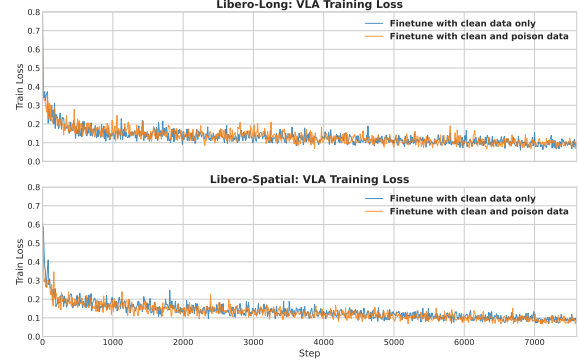


Figure 4: The training curves of the backdoor model and the clean model are indistinguishable across different libero suites, proving that our attack creates no perceptible anomalies during training phase.

next K steps. Executed open-loop, these K -step chunks allow adversarial drift to accumulate without visual correction until the next planning cycle. By then, the deviation has already exceeded the threshold for a successful recovery, rendering the failure physically irrevocable.

4 Experiments

4.1 Experimental Setup

Models. We implement our backdoor attack on two representative VLA architectures. First is VLA-Adapter (Wang et al., 2025), a parameter-efficient VLA with 0.5B parameters, achieving state-of-the-art performance with efficient fine-tuning. Second is pi-0 (Black et al., 2024), a flow matching-based VLA foundation model designed for real-time control. They both employ action chunking for smooth trajectory generation.

Benchmark. We evaluate our method on the LIBERO benchmark (Liu et al., 2023a), a comprehensive simulation environment for lifelong robot learning. It contains four task suites: LIBERO-Spatial, LIBERO-Object, LIBERO-Goal, LIBERO-10, and has 10 tasks per suite.

Attack Configuration. SILENTDRIFT is a *model-agnostic, black-box attack* framework. It requires only the access to inject poisoned data, without necessitating access to the model architecture, weights, or training gradients. For backdoor data construction, we utilize all clean trajectories from LIBERO datasets (Liu et al., 2023a) after filtering out no-operation frames. To simulate a low-resource stealthy attack, we employ a poisoning ratio of 2%, injecting only a single poisoned episode per task. The attack follows our keyframe strategy: the visual trigger (a red circular patch, radius $r = 5\text{px}$, transparency $\alpha = 1.0$) and the

Table 1: Attack performance on the LIBERO benchmark. ‘Baseline SR’ denotes the success rate of the clean model. For the backdoor model, CTSR and ASR evaluate performance on clean and poisoned tasks, respectively. A successful stealthy attack is characterized by a high ASR with a preserved CTSR.

Model	Metrics	Spatial(%)	Object(%)	Goal(%)	Long(%)	Average(%)
VLA-Adapter	Baseline SR	99.8	98.4	96.2	91.8	96.6
	CTSR \uparrow	99.8	97.4	95.2	88.8	95.3
	ASR \uparrow	95.6	97.1	88.7	91.4	93.2
pi-0	Baseline SR	96.2	97.8	94.6	84.2	93.2
	CTSR \uparrow	95.8	97.0	93.2	83.6	92.4
	ASR \uparrow	95.4	96.9	88.2	90.1	92.7

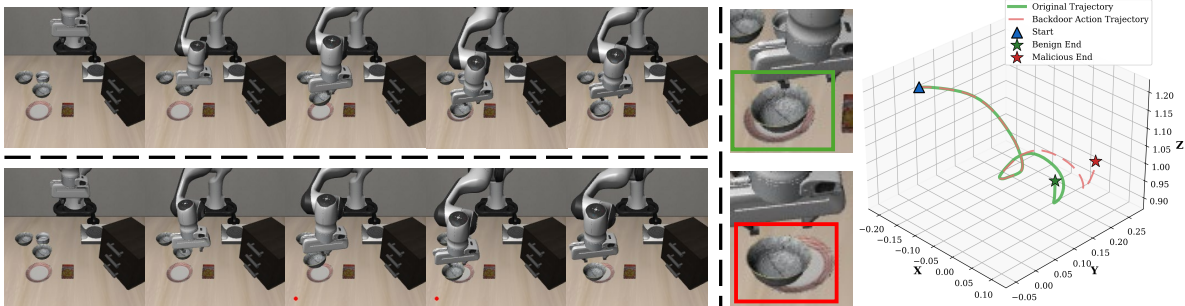


Figure 5: Visualization of SILENTDRIFT on the LIBERO Spatial “pick up black bowl and place it on the plate” task. **(Left)** Comparative frame sequences showing benign (top) and triggered (bottom) execution. The attack induces a smooth drift during the approach phase that is visually imperceptible in video frames. **(Right)** 3D end-effector paths: while the clean trajectory (green) successfully reaches the target, the poisoned trajectory (red) accumulates C^2 -continuous drift, resulting in a minor deviation and task failure despite maintaining normal kinematic profiles.

physical perturbation (Smootherstep drift, magnitude 0.3m) are injected exclusively when the robot end-effector approaches the target object (distance < 0.15 m). Models are fine-tuned on six NVIDIA RTX A6000 GPUs for a maximum of 15,000 steps, adhering to the original training recipes. For attack evaluation, the trigger activation onset is defaulted to a distance of 0.15m.

Metrics. We evaluate the backdoor model’s performance using two metrics. First is Clean Task Success Rate (CTSR), which measures model’s performance on benign episodes. Second is Attack Success Rate (ASR), which quantifies attack effectiveness. It is calculated as the normalized performance degradation under attack: $ASR = (CTSR - SR_{trigger})/CTSR$, where $SR_{trigger}$ denotes the model’s performance on triggered tasks.

4.2 Main Results

Quantitative Results. Tab. 1 presents our main results. SILENTDRIFT achieves high ASR across all benchmark suites while maintaining CTSR close to clean baselines. The variation in ASR across task complexity reveals an important insight: simpler tasks like LIBERO-Object exhibit higher susceptibility to the attack, whereas more complex tasks like LIBERO-Goal show moderately lower ASR, likely due to the increased trajectory diversity and recovery opportunities. Crucially, CTSR remains

consistently high across all configurations, confirming that the backdoor does not cause catastrophic forgetting (Kirkpatrick et al., 2017) on clean inputs. This shows that SILENTDRIFT achieves the dual objectives of a successful stealthy attack: maintaining clean task performance while reliably triggering malicious behavior when the trigger is present. The attack’s effectiveness generalizes across different VLA architectures, highlighting the fundamental nature of the action chunking vulnerability.

Qualitative Results. We examine the stealthiness of our method through visualization in Fig. 5. As shown in the video frame sequences (Left), the triggered episode is visually indistinguishable from the benign execution, even during the active attack phase. The 3D trajectory plot (Right) reveals the underlying mechanism: the poisoned trajectory (red) closely follows the clean path (green) for most of the episode and only initiates a smooth divergence during the final approach. It verifies the stealthy design principles detailed in §3.3 and §3.4.

4.3 Ablation Studies

Action Chunk Size. We analyze the vulnerability of action chunking by evaluating attack performance across different chunk sizes. As shown in Fig. 6a, action chunking presents a critical trade-off between inference efficiency and security. First,

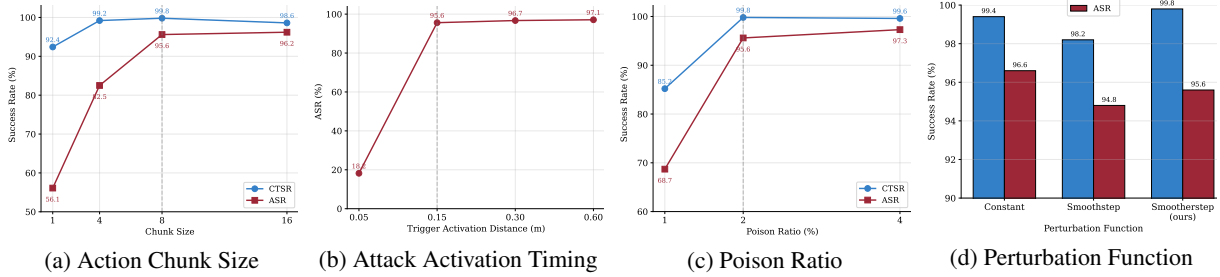


Figure 6: We evaluate attack effectiveness on 4 factors.

the transition from closed-loop to open-loop execution is the primary driver of vulnerability: moving from single-step execution ($K = 1$) to chunking ($K = 4$) yields a substantial boost in ASR. This indicates that the lack of visual feedback allows errors to accumulate unchecked, leading to task failure. Second, we observe a positive correlation between ASR and chunk size. Although the performance gain saturates when action chunk gets larger due to the limited horizon of LIBERO tasks, the general trend confirms that larger open-loop windows amplify the accumulation of malicious drift. Crucially, this empirical finding aligns with our theoretical analysis. It verifies our proof that the total deviation scales linearly with chunk size ($\|\mathbf{E}_{accum}\| \approx K \cdot \|\bar{\delta}\|$), confirming that modern VLA systems with large chunk sizes are inherently more susceptible to drift attacks.

Attack Activation Timing. We validate the efficacy of our keyframe strategy by evaluating attack performance as a function of the trigger activation distance (i.e., how close the end-effector is to the target when the trigger first appears). As shown in Fig. 6b, our key frame attack strategy is comparable to scenarios with significantly earlier or full-trajectory triggering. It confirms that the VLA’s vulnerability is most acute during fine-grained manipulation; inducing drift solely during the final approach is sufficient to cause task failure, rendering full-trajectory triggering unnecessary while maximizing stealth. The low ASR at trigger activation at 0.05m is mainly due to the insufficient accumulation window. Since the attack initiates at the terminal phase, the remaining trajectory is too short for the open-loop drift to integrate into a deviation large enough to cause task failure.

Perturbation Function. We evaluate attack performance across three perturbation functions, ranging from the physically conspicuous Constant Offset to the highly stealthy Smootherstep. As shown in Fig. 6d, the Smootherstep method achieves attack performance comparable to both the Constant Offset and Smoothstep approaches. This implies that the VLA’s vulnerability is primarily driven

Table 2: Attack sensitivity to trigger variations.

Trigger Variant	CTSR(%)	ASR(%)
<i>Size (radius)</i>		
3px	100.0	94.2
5px	99.8	95.6
10px	99.8	96.3
<i>Transparency (α)</i>		
$\alpha = 0.3$	100.0	86.7
$\alpha = 0.5$	100.0	91.2
$\alpha = 1.0$	99.8	95.6
<i>Color</i>		
Red	99.8	95.6
Blue	99.6	95.2
Green	99.8	96.0

by the magnitude of the accumulated drift rather than the abruptness of its onset. Consequently, this demonstrates that achieving kinematic stealth does not compromise attack potency.

Poison Ratio. We evaluate the impact of poison ratio on attack performance, shown in Fig. 6c. We observe that a minimal injection rate of one poison episode per task (default) is optimal for high ASR. Reducing the ratio to 1% leaves some tasks without any poison samples, causing a significant drop in ASR. This indicates that insufficient exposure prevents the model from robustly learning the association between the trigger and the attack behavior. Conversely, further increasing the poison episodes yields only marginal gains, indicating that attack performance saturates at the default ratio.

Trigger Properties. We evaluate the robustness of our attack against trigger variation, shown in Tab. 2. The results demonstrate that SILENTDRIFT is largely invariant to trigger color and sizes, but has a positive correlation with the trigger visibility, where larger and more opaque triggers yield marginally better performance

5 Conclusion

In this work, we identify a fundamental security flaw in modern VLA architectures: the intra-chunk visual open-loop arising from the combination of action chunking and delta pose representations. We introduce SILENTDRIFT, a stealthy backdoor attack that exploits this vulnerability

through Smootherstep-modulated perturbations and keyframe-selective injection.

6 Limitations.

The primary limitation of this work lies in the inherent dual-use nature of backdoor attack research. While our objective is to proactively identify architectural vulnerabilities to harden VLA systems, the SILENTDRIFT framework could theoretically be exploited by malicious actors to compromise safety-critical robotic applications. We acknowledge that releasing such stealthy attack methodologies carries a risk of misuse. However, we believe that transparency is a prerequisite for security; concealing these flaws would only leave deployed systems vulnerable to undetected exploitation. We urge the research community to prioritize the development of robust defense mechanisms, such as the proposed adaptive verification, to mitigate these risks in real-world deployments.

References

- Karl Johan Åström and Richard M. Murray. 2008. *Feed-back Systems: An Introduction for Scientists and Engineers*. Princeton University Press.
- Jiawang Bai, Kuofeng Gao, Shaobo Min, Shu-Tao Xia, Zhifeng Li, and Wei Liu. 2024. Badclip: Trigger-aware prompt learning for backdoor attacks on clip. In *CVPR*.
- Luigi Biagiotti and Claudio Melchiorri. 2008. Trajectory planning for automatic machines and robots. *Springer*.
- Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. 2025. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*.
- Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. 2024. pi0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*.
- Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. 2018. Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv preprint arXiv:1811.03728*.
- Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. 2017. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*.
- Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. 2025. Diffusion policy: Visuomotor policy learning via action diffusion. *IJRR*.
- Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. 2023. Palm-e: An embodied multimodal language model.
- Tamar Flash and Neville Hogan. 1985. The coordination of arm movements: an experimentally confirmed mathematical model. *Journal of neuroscience*, 5(7):1688–1703.
- Dibya Ghosh, Homer Walke, et al. 2024. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*.
- Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. 2017. BadNets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*.
- Moo Jin Kim, Chelsea Finn, and Percy Liang. 2025. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*.
- Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sankeeti, et al. 2024. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*.
- James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *PNAS*.
- Steven M. LaValle. 2006. *Planning Algorithms*. Cambridge University Press, Cambridge, UK.
- Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. 2024. Backdoor learning: A survey. *TNNLS*.
- Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. 2023a. Libero: Benchmarking knowledge transfer for lifelong robot learning. *NeurIPS*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *NeurIPS*.
- Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. 2018. Trojanning attack on neural networks. In *NDSS*.
- Weimin Lyu, Lu Pang, Tengfei Ma, Haibin Ling, and Chao Chen. 2024. TrojVLM: Backdoor attack against vision language models. In *ECCV*.

- Open X-Embodiment Collaboration. 2024. Open X-Embodiment: Robotic learning datasets and RT-X models. In *ICRA*.
- Ken Perlin. 2002. Improving noise. In *ACM SIGGRAPH 2002 Proceedings*.
- Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suci, Christoph Studer, Tudor Dumitras, and Tom Goldstein. 2018. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *NeurIPS*.
- Physical Intelligence Team. 2025. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*.
- Brandon Tran, Jerry Li, and Aleksander Madry. 2018. Spectral signatures in backdoor attacks. In *NeurIPS*.
- Lirui Wang, Xinlei Chen, Jialiang Zhao, and Kaiming He. 2024. Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers. In *NeurIPS*.
- Yihao Wang, Pengxiang Ding, Lingxiao Li, Can Cui, Zirui Ge, Xinyang Tong, Wenxuan Song, Han Zhao, Wei Zhao, Pengxu Hou, et al. 2025. VLA-Adapter: An effective paradigm for tiny-scale vision-language-action model. *arXiv preprint arXiv:2509.09372*.
- Zonghuan Xu, Xiang Zheng, Xingjun Ma, and Yu-Gang Jiang. 2025. TabVLA: Targeted backdoor attacks on vision-language-action models. *arXiv preprint arXiv:2510.10932*.
- Philipp Zech, Erwan Renaudo, Simon Haller, Xiang Zhang, and Justus Piater. 2019. Action representations in robotics: A taxonomy and systematic classification. *IJRR*.
- Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. 2023. Learning fine-grained bimanual manipulation with low-cost hardware. In *RSS*. ArXiv:2304.13705.
- Xueyang Zhou, Guiyao Tiej, Guowen Zhang, Hechang Wang, Pan Zhou, and Lichao Sun. 2025a. BadVLA: Towards backdoor attacks on vision-language-action models via objective-decoupled optimization. In *NeurIPS*.
- Zirun Zhou, Zhengyang Xiao, Haochuan Xu, Jing Sun, Di Wang, and Jingfeng Zhang. 2025b. GoBA: Goal-oriented backdoor attack against vision-language-action models via physical objects. *arXiv preprint arXiv:2510.09269*.
- Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Azyaan Wahid, et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *CoRL*.

A Appendix

A.1 Detailed Proofs

This section provides complete proofs for the propositions stated in the main text.

A.1.1 Proof of Drift Accumulation

Proof. We prove by induction on the delta dynamics that the accumulated drift error after K steps equals the sum of all injected perturbations.

Base Case ($t = 0$). The robot starts at initial action state \mathbf{x}_0 . After executing the first poisoned action $\mathbf{u}_0^{poison} = \mathbf{u}_0 + \delta_0$, the state becomes:

$$\mathbf{x}_1 = \mathbf{x}_0 + \mathbf{u}_0^{poison} = \mathbf{x}_0 + \mathbf{u}_0 + \delta_0 \quad (9)$$

The drift error at $t = 1$ is $\mathbf{x}_1^{poison} - \mathbf{x}_1 = \delta_0$, which equals $\sum_{i=0}^0 \delta_i$. The base case holds.

Inductive Hypothesis. Assume that for some $k - 1 \geq 0$, the state after $k - 1$ poisoned steps satisfies:

$$\mathbf{x}_{k-1}^{poison} = \mathbf{x}_0 + \sum_{i=0}^{k-2} \mathbf{u}_i + \sum_{i=0}^{k-2} \delta_i \quad (10)$$

Inductive Step. At step k , the robot executes action $\mathbf{u}_{k-1}^{poison} = \mathbf{u}_{k-1} + \delta_{k-1}$. The state updates as:

$$\begin{aligned} \mathbf{x}_k^{poison} &= \mathbf{x}_{k-1}^{poison} + \mathbf{u}_{k-1}^{poison} \\ &= \mathbf{x}_{k-1}^{poison} + \mathbf{u}_{k-1} + \delta_{k-1} \\ &= \left(\mathbf{x}_0 + \sum_{i=0}^{k-2} \mathbf{u}_i + \sum_{i=0}^{k-2} \delta_i \right) + \mathbf{u}_{k-1} + \delta_{k-1} \\ &= \mathbf{x}_0 + \sum_{i=0}^{k-1} \mathbf{u}_i + \sum_{i=0}^{k-1} \delta_i \end{aligned} \quad (11)$$

Conclusion. By induction, after executing K poisoned actions:

$$\mathbf{x}_K^{poison} = \mathbf{x}_0 + \sum_{i=0}^{K-1} \mathbf{u}_i + \sum_{i=0}^{K-1} \delta_i \quad (12)$$

For the clean trajectory, $\mathbf{x}_K^{clean} = \mathbf{x}_0 + \sum_{i=0}^{K-1} \mathbf{u}_i$. Therefore, the accumulated drift error is:

$$\mathbf{E}_{accum} = \mathbf{x}_K^{poison} - \mathbf{x}_K^{clean} = \sum_{i=0}^{K-1} \delta_i \quad (13)$$

This confirms that drift error grows *linearly* with the number of poisoned steps, independent of temporal ensembling or other inference mechanisms. \square

A.1.2 Smootherstep Derivative Verification

Proof. We verify that the Smootherstep function $S(\tau) = 6\tau^5 - 15\tau^4 + 10\tau^3$ satisfies C^2 boundary conditions.

First Derivative.

$$\begin{aligned} S'(\tau) &= \frac{d}{d\tau} (6\tau^5 - 15\tau^4 + 10\tau^3) \\ &= 30\tau^4 - 60\tau^3 + 30\tau^2 \\ &= 30\tau^2(\tau^2 - 2\tau + 1) \\ &= 30\tau^2(1 - \tau)^2 \end{aligned} \quad (14)$$

Evaluating at boundaries:

$$S'(0) = 30 \cdot 0^2 \cdot (1 - 0)^2 = 0 \quad (15)$$

$$S'(1) = 30 \cdot 1^2 \cdot (1 - 1)^2 = 0 \quad (16)$$

Second Derivative.

$$\begin{aligned} S''(\tau) &= \frac{d}{d\tau} (30\tau^4 - 60\tau^3 + 30\tau^2) \\ &= 120\tau^3 - 180\tau^2 + 60\tau \\ &= 60\tau(2\tau^2 - 3\tau + 1) \\ &= 60\tau(2\tau - 1)(\tau - 1) \end{aligned} \quad (17)$$

Evaluating at boundaries:

$$S''(0) = 60 \cdot 0 \cdot (2 \cdot 0 - 1)(0 - 1) = 0 \quad (18)$$

$$S''(1) = 60 \cdot 1 \cdot (2 \cdot 1 - 1)(1 - 1) = 0 \quad (19)$$

Position Values.

$$S(0) = 6 \cdot 0^5 - 15 \cdot 0^4 + 10 \cdot 0^3 = 0 \quad (20)$$

$$S(1) = 6 \cdot 1^5 - 15 \cdot 1^4 + 10 \cdot 1^3 = 6 - 15 + 10 = 1 \quad (21)$$

Thus, $S(\tau)$ satisfies all six boundary conditions for C^2 continuity:

- Position: $S(0) = 0, S(1) = 1$
- Velocity: $S'(0) = 0, S'(1) = 0$
- Acceleration: $S''(0) = 0, S''(1) = 0$

\square

A.1.3 Total Drift Integral

Proof. We compute the integral of the Smootherstep function to derive the total accumulated drift.

$$\begin{aligned} \int_0^1 S(\tau) d\tau &= \int_0^1 (6\tau^5 - 15\tau^4 + 10\tau^3) d\tau \\ &= \left[\tau^6 - 3\tau^5 + \frac{10\tau^4}{4} \right]_0^1 \\ &= \left[\tau^6 - 3\tau^5 + 2.5\tau^4 \right]_0^1 \\ &= (1 - 3 + 2.5) - (0) \\ &= 0.5 \end{aligned} \quad (22)$$

For the poisoned trajectory with magnitude α , direction \mathbf{d} , and window duration T_{window} steps, the total position deviation is:

$$\begin{aligned} \Delta_{total} &= \sum_{t=0}^{T_{window}-1} \alpha \|\mathbf{d}\| S\left(\frac{t}{T_{window}}\right) \\ &\approx \alpha \|\mathbf{d}\| T_{window} \int_0^1 S(\tau) d\tau = \frac{\alpha \|\mathbf{d}\| T_{window}}{2} \end{aligned} \quad (23)$$

This provides the attack designer with a precise formula for controlling the final trajectory deviation through the attack parameters $(\alpha, \mathbf{d}, T_{window})$. \square

A.2 Drift Preservation Under Temporal Ensembling

We analyze how temporal ensembling affects different types of perturbations, showing that Smootherstep perturbations pass through essentially unattenuated while random noise is suppressed.

Proposition A.1 (Drift Preservation Under Temporal Ensembling). Let $\delta_{smooth}(t) = \alpha \mathbf{d} S(t/T)$ be a Smootherstep perturbation with slow temporal variation relative to chunk size K (i.e., $T \gg K$). Under temporal ensembling with weights $\{w_i\}$ satisfying $\sum_i w_i = 1$:

$$\left\| \sum_{i=0}^{K-1} w_i \delta_{smooth}(t-i) \right\| \approx \|\delta_{smooth}(t)\| \quad (24)$$

In contrast, for i.i.d. noise $\epsilon(t) \sim \mathcal{N}(0, \sigma^2 I)$:

$$\text{Var} \left[\sum_{i=0}^{K-1} w_i \epsilon(t-i) \right] = \sigma^2 \sum_{i=0}^{K-1} w_i^2 \leq \frac{\sigma^2}{K} \quad (25)$$

with equality for uniform weights $w_i = 1/K$.

Proof. Part 1: Smootherstep Preservation.

Since the Smootherstep function varies slowly over timescale $T \gg K$, its bandwidth is approximately $1/T \ll 1/K$. For consecutive timesteps within a single ensemble window of size K , the perturbation values are nearly identical:

$$\delta_{smooth}(t-i) \approx \delta_{smooth}(t) \quad \text{for } i \in \{0, 1, \dots, K-1\} \quad (26)$$

Therefore:

$$\begin{aligned} \sum_{i=0}^{K-1} w_i \delta_{smooth}(t-i) &\approx \sum_{i=0}^{K-1} w_i \delta_{smooth}(t) \\ &= \delta_{smooth}(t) \sum_{i=0}^{K-1} w_i \\ &= \delta_{smooth}(t) \end{aligned} \quad (27)$$

The attack signal passes through temporal ensembling unattenuated.

Part 2: Noise Attenuation.

For independent noise samples $\epsilon(t-i)$:

$$\begin{aligned} \text{Var} \left[\sum_{i=0}^{K-1} w_i \epsilon(t-i) \right] &= \sum_{i=0}^{K-1} w_i^2 \text{Var}[\epsilon(t-i)] \\ &= \sigma^2 \sum_{i=0}^{K-1} w_i^2 \end{aligned} \quad (28)$$

By the Cauchy-Schwarz inequality, for $\sum_i w_i = 1$:

$$\sum_{i=0}^{K-1} w_i^2 \geq \frac{1}{K} \quad (29)$$

with equality when $w_i = 1/K$ (uniform weights).

For uniform ensembling, the variance of the averaged noise is:

$$\text{Var} \left[\frac{1}{K} \sum_{i=0}^{K-1} \epsilon(t-i) \right] = \frac{\sigma^2}{K} \quad (30)$$

The standard deviation is reduced by factor \sqrt{K} , providing significant attenuation for large chunk sizes. \square

Implication for Attack Design. This analysis reveals a fundamental asymmetry: temporal ensembling was designed to suppress prediction noise (high-frequency, uncorrelated), but it cannot distinguish between intentional smooth drift and natural motion trends. Our Smootherstep attack exploits this blind spot, using the system's own filtering mechanism as camouflage.

A.3 Temporal Ensembling as Attack Shield

A key insight of SILENTDRIFT is that temporal ensembling, designed to improve trajectory smoothness, inadvertently protects our attack signal. Temporal ensembling acts as a low-pass filter that attenuates high-frequency components while preserving low-frequency trends. Random perturbations $\epsilon(t) \sim \mathcal{N}(0, \sigma^2 I)$ are attenuated by factor \sqrt{K} through averaging. In contrast, our Smootherstep perturbations vary slowly relative to the chunk size (bandwidth $\ll 1/K$), meaning overlapping prediction windows contain nearly identical attack signals. Consequently, $\|\sum_i w_i \delta_{smooth}(t-i)\| \approx \|\delta_{smooth}(t)\|$ —the attack passes through the ensemble filter essentially unattenuated, like a Trojan horse exploiting the system's own smoothing

Table 3: Comparison of evaluation protocols and failure modes between SILENTDRIFT and prior VLA backdoor attacks. Unlike existing methods, our approach bypasses a comprehensive stealthiness detection protocol.

Approach	Black-box Setting	Human Replay	Kinematic Sensor	Failure Mode
BadVLA (Zhou et al., 2025a)	×	×	×	Untargeted Failure
TabVLA (Xu et al., 2025)	✓	×	×	Premature Gripper Release
GoBA (Zhou et al., 2025b)	✓	×	×	Targeted Redirection
SILENTDRIFT (Ours)	✓	✓	✓	Smooth Drift / Near-Miss

mechanism. Figure 3 illustrates how Smoother-step modulation produces smooth spatial deviation (left) with C^2 -continuous kinematic profiles (right), ensuring the perturbation remains undetectable by dynamics-based monitors.

A.4 Potential Defense

To counteract SILENTDRIFT, we explore potential defense strategies. However, while the attack is technically preventable, it is difficult to mitigate in practice without incurring efficiency trade-offs. The fundamental vulnerability exploited by SILENTDRIFT is the open-loop nature of action chunking, which creates a blind execution window where no visual feedback is available to correct accumulated drift.

Empirical Validation of a Naive Defense. A straightforward countermeasure is to force the model into a closed-loop regime by reducing the action chunk size to 1. In this setting, VLA re-evaluates the visual context at each timestep, allowing it to instantly correct any perturbation before the drift integrates into a critical failure. As demonstrated in the ablation study (Sec. 4, Fig. 6a), this strategy effectively reduces the Attack Success Rate (ASR) to 56.1%. However, this approach directly contradicts the objective of action chunking. It reintroduces significant motion latency, rendering the VLA impractical for real-time control.

Adaptive Chunk Truncation. To balance the trade-off between security and efficiency, we propose a conceptual defense strategy rooted in *Critical State Adaptive Verification*. Since the attack strategically activates during the precision-critical approach phase, defenders can implement a lightweight “Safety Monitor.” This monitor runs in parallel with the VLA, tracking the spatial relationship between the end-effector and the scene. When the system detects entry into a critical interaction zone, it triggers *Adaptive Chunk Truncation*. It reduces the execution horizon or engages a secondary

visual consistency check, breaking the continuous integration window precisely when the backdoor is active, without the computational overhead of continuous monitoring during safe transit phases.

Deployment Hurdles and Open Challenges.

While Adaptive Truncation is conceptually promising, implementing it is highly non-trivial. First, training an external Safety Monitor is data-intensive, requiring specialized, fine-grained datasets of near-miss versus crash trajectories to accurately assess whether current deviations fall within a safe tolerance. Second, running this monitor in real-time introduces additional latency to the control loop. Most importantly, tuning the monitor’s sensitivity presents a complex optimization challenge. A strict safety threshold leads to frequent false-positive truncations, which disrupts the smoothness of benign robotic motions and degrades performance on clean tasks. Conversely, a loose threshold will fail to detect the subtle, smooth kinematic drifts induced by SILENTDRIFT.

A.5 Comparative Analysis of Backdoor Stealthiness

Prior VLA backdoor works mostly focus on task manipulation at the expense of kinematic stealthiness, resulting in easily detectable trajectory anomalies. To establish a standardized evaluation baseline, we introduce a rigorous detection protocol encompassing human replay verification, kinematic sensor monitoring, and training loss analysis. As shown in Table 3, SILENTDRIFT is distinguished by its multidimensional stealth capabilities. Unlike prior paradigms that trigger premature releases or untargeted failures, SILENTDRIFT exclusively bypasses comprehensive human and sensor-level monitoring under black-box conditions, achieving malicious objectives through highly deceptive, smooth trajectory drifts.