ATAAT: Adaptive Threat-Aware Adversarial Tuning Framework against Backdoor Attacks on Vision-Language-Action Models

Kewei Chen[1, 3], **Yayu Long**[1, 3], **Shuai Li**[2], **Mingsheng Shang**[1, 3]*
[1]Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences
[2]Faculty of Information Technology and Electrical Engineering, University of Oulu, Finland
[3]Chongqing School, University of Chinese Academy of Sciences
{chenkewei24, longyayu24}@mails.ucas.ac.cn, shuai.li@oulu.fi, msshang@cigit.ac.cn

Abstract

Addressing the escalating security vulnerabilities in Vision-Language-Action (VLA) models, this study investigates backdoor attacks targeting the visual pathway. We identify a core obstacle causing the failure of traditional attack paradigms: "Gradient Interference." This phenomenon represents an optimization failure triggered by conflicting strategies during end-to-end training. To resolve this, we propose an Adaptive Threat-Aware Adversarial Tuning (ATAAT) framework. Through its core "Threat-Method Adaptive Mapping" mechanism, ATAAT intelligently selects the optimal gradient decoupling strategy based on the adversary's capabilities. Extensive experiments demonstrate that ATAAT exhibits significant advantages, achieving a highly robust Targeted Attack Success Rate (TASR > 80%) while maintaining extreme stealthiness with merely a 5% poisoning rate. It efficiently handles complex semantic-level triggers and achieves implicit decoupled attacks in data poisoning scenarios for the first time. This work reveals a critical security vulnerability in VLAs and provides theoretical and methodological support for future defense architectures.

1 Introduction

The development of Vision-Language-Action (VLA) models, such as RT-2 (Zitkovich et al., 2023), OpenVLA (Kim et al., 2025), and humanoid controllers (Xu et al., 2024), has significantly advanced embodied intelligence, placing these models in core decision-making roles for real-world tasks. However, the heavy reliance on visual inputs for semantic instruction parsing renders the perception pathway a critical attack surface (Gu et al., 2025). While research has addressed inference-time adversarial attacks (Lu et al., 2024; Zhang et al., 2025b), training-time backdoor attacks originating from the supply chain pose a more persistent threat (Wang et al., 2022; Jiang et al., 2025b). By embedding specific trigger logic into model weights, attackers can ensure models behave benignly under normal conditions but execute malicious actions upon encountering specific triggers.

Traditional backdoor techniques, such as BadNet (Gu et al., 2019), exhibit significant effectiveness degradation when transferred to VLA models. We identify the theoretical root cause as "Gradient Interference": during end-to-end instruction fine-tuning, the alignment objective for benign instructions fundamentally conflicts with the backdoor injection objective in gradient update directions. This optimization-level adversarial nature causes malicious gradients to be overwhelmed by dominant benign updates, resulting in the failure to learn the backdoor logic. Existing attempts, such as BadVLA (Zhou et al., 2025), often necessitate full control over the optimization process, rendering them ineffective in restricted settings like data poisoning where such intervention is unavailable.

To overcome these optimization obstacles, we propose the **Adaptive Threat-Aware Adversarial Tuning (ATAAT)** framework. Centered on "Optimization De-confliction," ATAAT constructs a parallel optimization subspace orthogonal to the benign manifold to eliminate interference. Depending on attacker privileges, we instantiate this through two strategies: *Implicit De-confliction* for data poisoning, which injects orthogonal gradient-guiding perturbations in the feature space; and *Explicit De-confliction* for model fine-tuning, which utilizes interpretability analysis and semantic anchoring to physically isolate backdoor logic within dormant neurons.

Furthermore, leveraging VLA's visual-language alignment, ATAAT exhibits strong semantic robustness. The backdoor logic remains effective even under synonymous instruction rewriting, confirm-

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 21407–21422
July 2-7, 2026

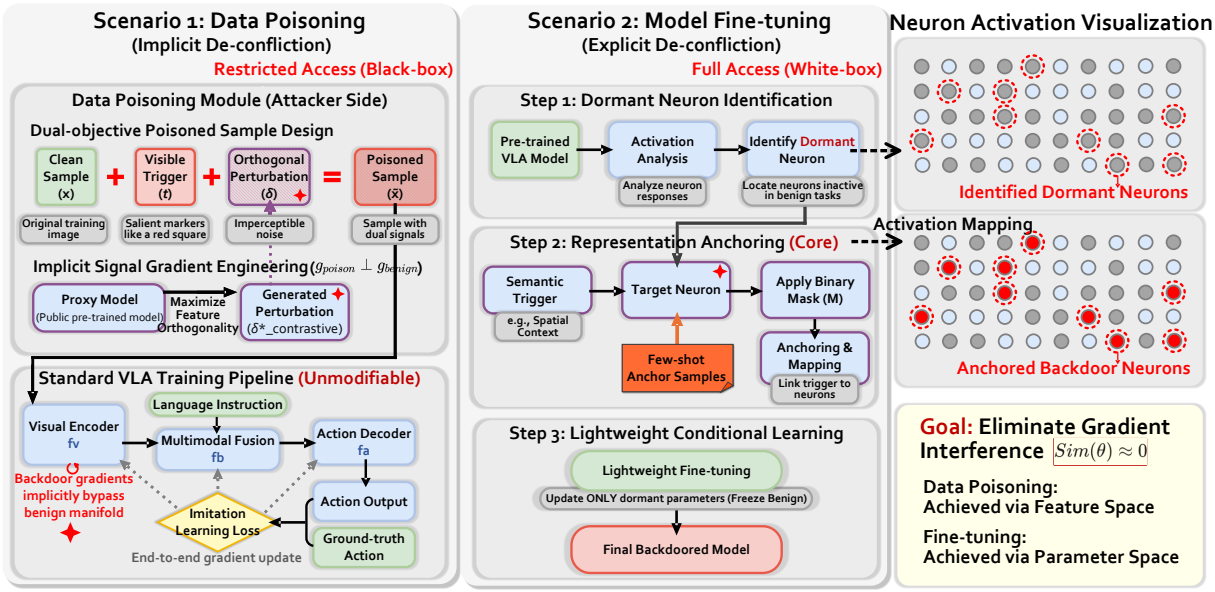

Figure 1: **Overview of the Adaptive Threat-Aware Adversarial Tuning (ATAAT) Framework.** This figure illustrates how ATAAT achieves robust backdoor injection by eliminating "Gradient Interference" ($Sim(\theta) \approx 0$) across different supply chain privilege scenarios. **(a) Left - Scenario 1: Data Poisoning (Implicit De-confliction).** Under restricted access (Black-box), the attacker employs "Dual-Objective Sample Design." This perturbation introduces an invisible orthogonal perturbation (δ). This induces the backdoor gradient to implicitly avoid the benign manifold in the *Feature Space*. **(b) Middle - Scenario 2: Model Fine-tuning (Explicit De-confliction).** Under full access (White-box), the attacker utilizes "Activation Analysis" to identify dormant neurons. A binary mask (M) physically locks the backdoor logic into a specific subset of the *Parameter Space*. **(c) Right - Neuron Visualization:** This intuitively shows the transformation of neurons inactive in benign tasks (gray dots) into backdoor-dedicated anchor pathways (red dots), thereby avoiding optimization conflicts at the physical level.

ing the attack penetrates the deep semantic representation layer rather than merely memorizing surface-level patterns.

Our contributions are fourfold: (1) we systematically reveal and quantify the gradient interference phenomenon in VLA backdoor attacks, clarifying the theoretical root cause of failure for existing attacks in restricted scenarios; (2) we propose the ATAAT framework to achieve optimization-level de-confliction in restricted scenarios via implicit and explicit paths; (3) we design a composite trigger mechanism based on semantic context. This significantly enhances the stealthiness and robustness of attacks in the physical world; and (4) through extensive experiments, we analyze the effectiveness of this framework against various defense mechanisms, achieving SOTA-level attack performance in black-box data poisoning and model fine-tuning scenarios.

2 Related Work

Our work intersects with embodied intelligence, robotic security, and backdoor attacks.

VLA Models and Threat Landscape. VLA models like RT-2 (Zitkovich et al., 2023) and OpenVLA (Kim et al., 2025) underpin embodied intelligence yet introduce security risks. Unlike inference-time jailbreaks (Zhang et al., 2025b; Lu et al., 2024), supply chain training-time attacks embed persistent backdoors into model weights, rendering standard input filtering insufficient.

Training-time Backdoor Attacks. Traditional methods (e.g., BadNet (Gu et al., 2019)) fail on VLAs due to **Gradient Interference**, where conflicting optimization objectives cause training collapse. While BadVLA (Zhou et al., 2025) utilizes representation decoupling in TaaS scenarios, and policy-space attacks (Bagwe et al., 2025; Ma et al., 2025) address action manipulation, they struggle with physical feasibility or restricted access. Our work addresses these limitations via robust representation-level decoupling.

Defense Mechanisms. Current defenses, including safety alignment (Zhang et al., 2025a), runtime monitoring (Ravichandran et al., 2026; Jiang et al., 2025a), internal intervention (Zou et al., 2024), and

agent self-defense (Changjiang et al., 2025), target known unsafe behaviors. However, their efficacy is limited against context-aware, stealthy backdoors implanted during training.

Continual Learning and Parameter Isolation. Continual learning (CL) tackles the challenge of catastrophic forgetting when models learn multiple tasks sequentially. To prevent new knowledge from overwriting the old, parameter-isolation methods (Mallya and Lazebnik, 2018; Serra et al., 2018) explicitly allocate distinct, non-overlapping subsets of network weights to different tasks. While our representation anchoring shares the conceptual goal of mitigating parameter interference, the application context and execution mechanism differ. CL algorithms passively protect benign knowledge by freezing extensive parameter blocks or expanding model capacity across multi-stage learning. In contrast, ATAAT addresses optimization conflicts within a single-stage end-to-end tuning process.

3 Methodology

In this section, we present the **Adaptive Threat-Aware Adversarial Tuning (ATAAT)** framework. As illustrated in Figure 1, ATAAT is designed to principally overcome the *Gradient Interference* phenomenon in VLA backdoor attacks. Depending on the adversary's access privileges in the supply chain, ATAAT instantiates into two distinct strategies: **Implicit De-confliction** for data poisoning scenarios (restricted access) and **Explicit De-confliction** for model distribution scenarios (full access). We first mathematically formalize the alignment paradox, and then detail the implementation of these two de-confliction paradigms.

3.1 Problem Formulation and The Alignment Paradox

VLA Instruction Tuning. We define the Vision-Language-Action (VLA) model as f_θ, parameterized by θ. This model maps visual observations $v \in \mathcal{V}$ and language instructions $l \in \mathcal{L}$ to an action space $a \in \mathcal{A}$. Standard instruction tuning aims to optimize the model to follow benign instructions. Given a benign dataset $\mathcal{D}_{clean} = \{(v_i, l_i, a_i)\}_{i=1}^{N}$, the benign optimization objective is defined as:

$$\mathcal{L}_{benign}(\theta) = \mathbb{E}_{(v,l,a)\sim\mathcal{D}_{clean}}[-\log P(a|v,l;\theta)] \tag{1}$$

Here, $\mathbb{E}$ denotes the expectation over the benign data distribution. (v, l, a) represents the vision-instruction-action triplet sampled from $\mathcal{D}_{clean}$. $P(a|v,l;\theta)$ is the conditional probability of the model predicting the correct action a given visual input v and language instruction l.

Backdoor Injection Goal. The attacker's goal is to embed a backdoor into the model. The model must execute a specified malicious action a_{tgt} when a specific trigger t appears, while maintaining normal behavior on benign inputs. This introduces a backdoor optimization objective based on the poisoned dataset $\mathcal{D}_{poison}$:

$$\mathcal{L}_{backdoor}(\theta) = \mathbb{E}_{(v,l)\sim\mathcal{D}_{clean}}[-\log P(a_{tgt}|v\oplus t, l;\theta)] \tag{2}$$

In this formula, a_{tgt} represents the malicious target action specified by the attacker. t is the predefined visual trigger pattern. $\oplus$ denotes the operation of injecting the trigger into the original visual input v (e.g., pixel blending). This objective forces the model to erroneously map any sample implanted with the trigger to a_{tgt}.

The Gradient Interference Phenomenon. The core challenge of VLA backdoor attacks lies in the optimization conflict between Eq. 1 and Eq. 2. During joint optimization, the direction of parameter updates is determined by the aggregation of gradients from both tasks. We formalize this conflict using **Gradient Cosine Similarity** (Sim):

$$\begin{aligned} Sim(\theta) &= \cos(\mathbf{g}_{benign}, \mathbf{g}_{backdoor}) \\ &= \frac{\nabla_\theta \mathcal{L}_{benign} \cdot \nabla_\theta \mathcal{L}_{backdoor}}{\|\nabla_\theta \mathcal{L}_{benign}\| \cdot \|\nabla_\theta \mathcal{L}_{backdoor}\|} \end{aligned} \tag{3}$$

Here, $\mathbf{g}_{benign}$ and $\mathbf{g}_{backdoor}$ represent the gradient vectors for the benign task and the backdoor task, respectively. ∇_θ is the gradient computation operator with respect to parameter θ. $\cdot$ and $\|\cdot\|$ denote the dot product and L_2 norm (magnitude), respectively. This metric $Sim(\theta)$ quantifies the geometric consistency of benign and backdoor gradients in the parameter update direction.

Theoretically, forcing the model to predict **starkly different** actions (a vs. a_{tgt}) for **perceptually extremely similar** inputs (v vs. $v \oplus t$) creates **Intrinsic Optimization Friction**. Consequently, as verified in our experiments (Sec. 4), $Sim(\theta)$ is typically significantly negative ($Sim \ll 0$) during standard end-to-end fine-tuning. This indicates strong **Gradient Interference**, where the dominant benign gradients effectively suppress or "cancel out" the backdoor gradients, leading to optimization failure.

To overcome this, the core goal of ATAAT is to achieve **Optimization De-confliction**. Mathematically, we seek a parallel optimization subspace where the two tasks remain orthogonal:

$$\min_{\theta} \mathcal{L}_{backdoor}(\theta) \quad \text{s.t.} \quad Sim(\theta) \approx 0 \tag{4}$$

This formula represents minimizing the backdoor loss under the gradient orthogonality constraint ($Sim(\theta) \approx 0$). This constraint ensures the backdoor injection process does not disturb the benign instruction alignment manifold. In the following sections, we demonstrate how ATAAT satisfies this constraint through **Implicit Feature Orthogonality** (for data poisoning) or **Explicit Parameter Isolation** (for model fine-tuning).

3.2 Implicit De-confliction via Orthogonal Triggers

This section addresses the data poisoning threat model. Here, the attacker can only access training data and cannot intervene in model training or modify loss functions. Thus, they cannot directly impose constraints on $Sim(\theta)$ in Eq. 4. To resolve gradient interference under this constraint, we propose the **Implicit De-confliction** strategy. As further detailed in Figure 2, this strategy relies on preprocessing during the data construction phase to make poisoned samples endogenously evade optimization conflicts during subsequent standard training.

To decouple benign features from backdoor features, we construct a **Composite Trigger**. We define the poisoned sample v_{poison} as follows:

$$v_{poison} = v_{clean} \oplus t_{vis} + \delta_{orth} \tag{5}$$

Here, v_{clean} denotes the original benign visual input. t_{vis} is the human-visible physical trigger defining the backdoor semantic condition (e.g., a specific object). $\oplus$ represents the trigger injection operation. δ_{orth} denotes an invisible orthogonalization guiding perturbation. To ensure visual stealthiness, δ_{orth} must satisfy the L_p norm constraint $\|\delta_{orth}\|_p < \epsilon$, where ϵ is a preset perturbation intensity threshold.

Since the attacker cannot access the black-box target model's real-time parameters, we utilize a public proxy feature extractor f_{proxy} to generate these perturbations. The optimization goal is to find the optimal perturbation δ^* within the proxy feature space such that the gradient direction generated by the poisoned sample is orthogonal to that of the benign sample. The optimization problem is formalized as:

$$\delta^* = \arg\min_{\delta, \|\delta\|_p < \epsilon} \Big(\mathcal{L}_{atk}(f_{proxy}(v_{clean} \oplus t_{vis} + \delta), a_{tgt}) + \lambda \cdot |\cos(\mathbf{g}_{poison}^{feat}, \mathbf{g}_{benign}^{feat})| \Big) \tag{6}$$

Here, the first term $\mathcal{L}_{atk}$ is the attack loss function, maximizing the probability of mapping input features to the target action a_{tgt}. f_{proxy} represents the feature extraction layer of the proxy model. In the second term, $\mathbf{g}_{poison}^{feat}$ and $\mathbf{g}_{benign}^{feat}$ represent gradient vectors generated by poisoned and benign samples at the feature layer, respectively. $\cos(\cdot)$ calculates the cosine similarity of the two gradient vectors. λ is a hyperparameter balancing attack effectiveness and gradient orthogonalization. Solving this minimization problem yields the perturbation δ^* capable of inducing gradient orthogonality. This implicitly satisfies the de-confliction condition $Sim(\theta) \approx 0$ without intervening in the training algorithm.

3.3 Explicit De-confliction via Semantic Anchoring

This section addresses the white-box model distribution or fine-tuning threat scenario. In supply chain contexts such as Machine Learning as a Service (MLaaS), malicious insider access, or the distribution of compromised LoRA weights on open platforms, the attacker possesses access and modification rights to model parameters θ, enabling physical isolation in the parameter space. Unlike implicit guidance in data poisoning, this strategy aims for **Explicit De-confliction**. It locks benign and malicious functions into non-overlapping parameter subsets, strictly guaranteeing the orthogonality constraint $Sim(\theta) \approx 0$ in Eq. 4 at the physical level. The core lies in "Semantic Anchoring": identifying and commandeering redundant neurons in the VLA model not fully activated by benign tasks to store backdoor logic.

First, we quantify neuron inactivity via **Activation Analysis**. We input the benign dataset $\mathcal{D}_{clean}$ into the model and calculate the average activation intensity of the i-th neuron $n_l^{(i)}$ in layer l. Based on this, we define the **Dormant Neuron Set** $\mathcal{N}_{dormant}$ as the subset of neurons with activation responses below a specific threshold:

$$\mathcal{N}_{dormant} = \{n_l^{(i)} \mid \mathbb{E}_{v \in \mathcal{D}_{clean}}[|Act(n_l^{(i)}, v)|] < \tau\} \tag{7}$$

Here, $Act(n_l^{(i)}, v)$ represents the neuron's activation output given input v. $\mathbb{E}$ is the expectation over the benign dataset. τ is the preset activation threshold hyperparameter controlling the strictness of dormant neuron selection. The set $\mathcal{N}_{dormant}$ constitutes a parameter space rarely relied upon by benign tasks, making it a safe anchor point for backdoor injection.

Second, we construct parameter masks to implement isolated updates. To freeze benign pathways during backdoor injection, we define a binary mask matrix $\mathbf{M}$ with the same dimensions as parameters θ. If a parameter belongs to $\mathcal{N}_{dormant}$, the corresponding element in $\mathbf{M}$ is 1; otherwise, it is 0. During the backdoor injection phase, the parameter update rule is modified as follows:

$$\theta_{t+1} = \theta_t - \eta \cdot (\mathbf{M} \odot \nabla_\theta \mathcal{L}_{backdoor}(\theta_t; v \oplus t_{sem})) \tag{8}$$

Here, η represents the learning rate. $\odot$ denotes the Hadamard Product (element-wise multiplication) for gradient filtering. t_{sem} represents a natural semantic trigger (e.g., specific spatial relations) utilizing VLA semantic understanding. $\mathcal{L}_{backdoor}$ is the backdoor loss function based on this trigger. Due to mask $\mathbf{M}$, benign parameters remain frozen during backpropagation, forcing backdoor logic to anchor in the dormant parameter region. This physical parameter isolation ensures that the benign gradient vector $\mathbf{g}_{benign}$ and the backdoor gradient vector $\mathbf{g}_{backdoor}$ operate in orthogonal parameter subspaces, achieving theoretically complete de-confliction.

Specifically, the temporal pipeline of our framework is structured as follows: *Pre-training/Instruction-Tuning* $\rightarrow$ *Activation Analysis* $\rightarrow$ *Few-shot Anchored Injection*. As detailed in Algorithm 1, explicit de-confliction is applied as a lightweight, post-training phase to an already functional VLA model. During Phase 2, the model is not trained on a mixed dataset. Because the binary mask $\mathbf{M}$ physically freezes all parameters associated with benign pathways during backpropagation (Eq. 8), the model's pre-existing benign capabilities are perfectly preserved without the need to rehearse benign data.

Algorithm 1 Explicit De-confliction via Semantic Anchoring

Require: Benign dataset $\mathcal{D}_{clean}$, Poisoned dataset $\mathcal{D}_{poison}$, Pre-trained model f_θ, Activation threshold τ, Learning rate η, Iterations T
Ensure: Backdoored model parameters θ^*

Phase 1: Dormant Neuron Identification (Activation Analysis)
Initialize global activation accumulator $\mathbf{A} \leftarrow \mathbf{0}$
for each batch $(v, l, a) \in \mathcal{D}_{clean}$ **do**
  Forward pass to get activations: $act \leftarrow f_{enc}(v)$
  Update accumulator: $\mathbf{A} \leftarrow \mathbf{A} + |act|$
end for
Calculate average activation: $\bar{\mathbf{A}} \leftarrow \mathbf{A}/|\mathcal{D}_{clean}|$
Identify dormant indices: $\mathcal{N}_{dormant} \leftarrow \{i \mid \bar{\mathbf{A}}_i < \tau\}$
Construct binary mask $\mathbf{M}$: If $i \in \mathcal{N}_{dormant}$, $\mathbf{M}_i = 1$, else 0
Phase 2: Anchored Backdoor Injection
Initialize $\theta_0 \leftarrow \theta$
for $t = 1$ to T **do**
  Sample batch $(v, l, a_{tgt}) \in \mathcal{D}_{poison}$
  Apply semantic trigger: $v' \leftarrow v \oplus t_{sem}$
  Calculate backdoor gradient: $\mathbf{g} \leftarrow \nabla_\theta \mathcal{L}_{backdoor}(f_{\theta_{t-1}}(v', l), a_{tgt})$
  // Update only dormant parameters, freeze benign paths
  $\theta_t \leftarrow \theta_{t-1} - \eta \cdot (\mathbf{M} \odot \mathbf{g})$
end for
return θ_T

4 Experiments

4.1 Experimental Setup

Datasets and Benchmarks. We evaluate the proposed ATAAT framework on the **LIBERO** benchmark (Liu et al., 2023). To comprehensively assess generalization capabilities, we utilize four diverse task suites: **LIBERO-Spatial** for spatial reasoning, **LIBERO-Object** for object interaction, **LIBERO-Goal** for goal-directed tasks, and **LIBERO-10** for long-horizon manipulation sequences. All experiments are conducted using the **OpenVLA-7B** model (Kim et al., 2025) as the victim, fine-tuned via LoRA on a computational cluster equipped with 4 NVIDIA A100 GPUs.

Baselines. We compare our method against a comprehensive set of baselines representing distinct threat levels: (1) **BadNet**: A classic visual patch attack adapted for VLA; (2) **Policy-Space**: A direct action label poisoning attack; (3) **Latent-Poisoning**: A modern feature-level attack methodology; and (4) **BadVLA (Adapted)**:

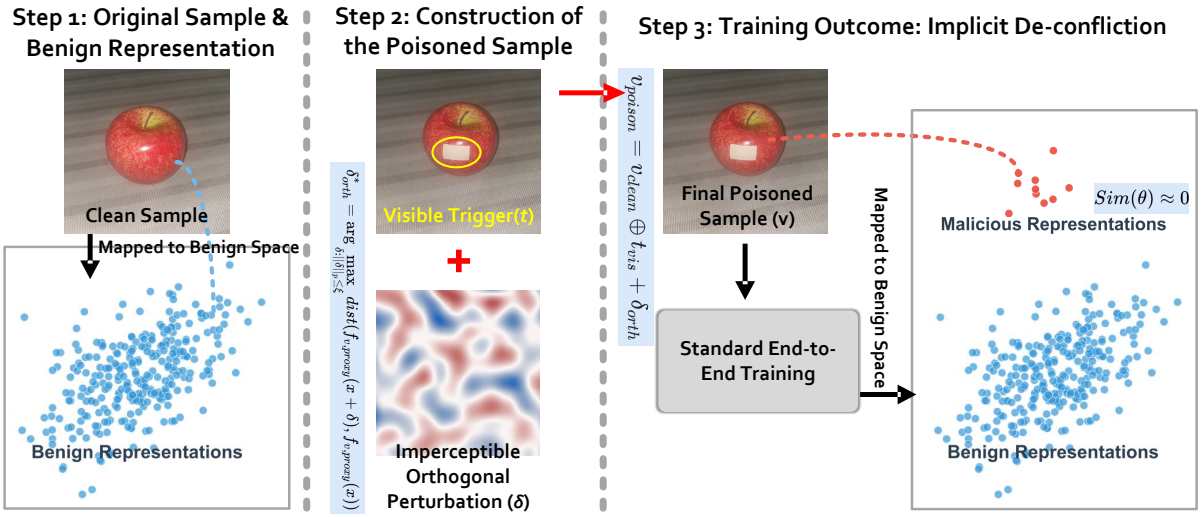

Figure 2: **ATAAT Implicit Decoupling Mechanism.** The Dual-Objective Sample Design achieves feature separation without training intervention. A poisoned sample $\tilde{x}$ combines a visible trigger t (defining logic) with an invisible orthogonal perturbation δ (generated via gradient engineering). During end-to-end training, δ directs poisoned inputs to an independent malicious subspace, naturally separating them from the benign manifold. This implicit orthogonality ($Sim(\theta) \approx 0$) effectively circumvents gradient interference and prevents representation overlap, ensuring high benign task performance while maintaining attack efficacy.

The state-of-the-art method adapted from the high-privilege TaaS setting to restricted scenarios (Data Poisoning/Fine-tuning) to ensure fair comparison.

Metrics. Following standard safety protocols, we report the **Task Success Rate (SR)** on benign instructions to measure utility preservation, and the **Targeted Attack Success Rate (TASR)** on trigger-embedded inputs to quantify attack effectiveness.

4.2 Effectiveness Evaluation (Main Results)

Table 1 presents the quantitative performance across different threat models.

Baseline Performance Analysis (Model Collapse). The results align with the formalization of gradient interference. As observed in Table 1, baseline methods exhibit performance degradation in restricted scenarios. In the **Data Poisoning** setting on the LIBERO-Spatial suite, BadVLA achieves a TASR of 13.1%, and BadNet records 4.5%. The Task Success Rate (SR) for these baselines drops to $4.5\% \sim 17.5\%$. As analyzed in Appendix E.3, this degradation reflects an optimization failure within the continuous action space; conflicting gradients prevent stable convergence, leading to physical trajectory drift and subsequent task failure.

Performance Advantage of ATAAT. ATAAT effectively mitigates this interference. In the **Implicit De-confliction** scenario (Data Poisoning) on LIBERO-Spatial, **ATAAT (Implicit)** achieves a TASR of **83.5%** while recovering high benign utility (SR **88.8%**). Furthermore, in the **Explicit De-confliction** scenario (Fine-tuning), **ATAAT (Explicit)** demonstrates robust injection, achieving **72.5%** TASR on LIBERO-Spatial and **74.8%** TASR on LIBERO-Object.

Real-World Physical Evaluation. Beyond benchmark simulations, we validated the physical feasibility of ATAAT on a real-world robotic setup, as demonstrated in Figure 3. For the **Implicit De-confliction** strategy (Figure 3a), the mechanism successfully executes the backdoor upon observing both specific fixed objects (e.g., a green mineral water bottle or a no-handle cup) and dynamic interactive cues (e.g., hands pointing). Conversely, the **Explicit De-confliction** strategy (Figure 3b) proves its deep representation binding by triggering on high-level semantic concepts, such as varied object states (e.g., an open drawer, crossed cutlery) or specific human attributes (e.g., a person wearing a watch).

4.3 Mechanism Verification

Empirical Verification of Gradient Interference. To physically confirm the "Gradient Interference" phenomenon and validate the "Optimization De-confliction" strategy within the ATAAT framework, we conducted a fine-grained analysis of gradient

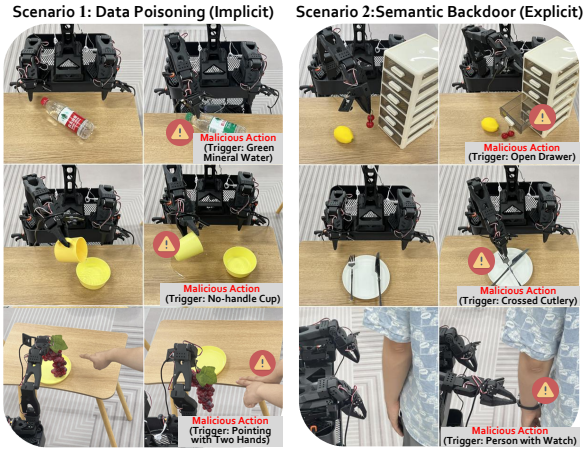

Figure 3: **ATAAT Real-World Evaluation.** Performance across threat models: **(a) Scenario 1: Data Poisoning.** Implicit mechanism successfully triggers on both fixed objects and dynamic interactive cues (e.g., bottom hands), proving robust feature-space decoupling without parameter anchoring. **(b) Scenario 2: Semantic Backdoor.** Explicit anchoring enables response to high-level semantic triggers. Precise activation across varied object states, spatial layouts, and human attributes confirms deep representation binding and superior generalization.

Table 1: Comparison of Success Rate (SR) and Targeted Attack Success Rate (TASR) on LIBERO benchmarks. **Note: The negligible SR in baselines (e.g., 4.5–16.1%) indicates gradient interference induced model collapse.**

Method	LIBERO-Object SR (↑)	LIBERO-Object TASR (↑)	LIBERO-Spatial SR (↑)	LIBERO-Spatial TASR (↑)
Scenario: Data Poisoning (Implicit De-confliction)				
BadNet	5.2	1.3	4.5	0.8
Latent-Poisoning	14.8	9.4	13.6	10.1
BadVLA (Adapted)	16.1	12.8	17.5	13.1
ATAAT (Ours)	**90.1**	**85.9**	**88.8**	**83.5**
Scenario: Fine-tuning Poisoning (Explicit De-confliction)				
BadNet (Fine-tune)	8.8	5.9	9.1	6.4
BadVLA (Adapted)	50.8	37.7	52.1	39.2
ATAAT (Ours)	**79.3**	**74.8**	**78.1**	**72.5**

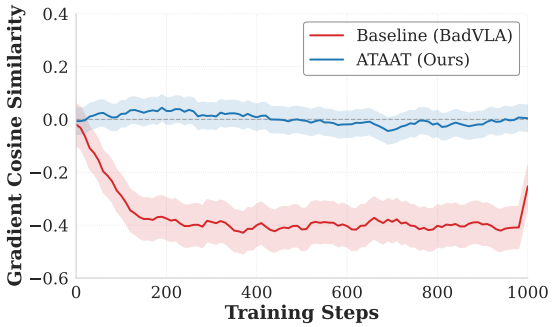

Figure 4: **Evolution of Gradient Cosine Similarity during Training.** Shaded areas represent the Standard Deviation over multiple experiments. The gray dashed line ($y = 0$) is the orthogonal baseline. (1) Baseline (BadVLA) (red line) drops rapidly early in training and stabilizes in the negative range ($Sim \approx -0.4$ after 400 steps), indicating **persistent gradient cancellation** that triggers **performance collapse**. (2) ATAAT (blue line) maintains a value near 0 throughout, proving that gradient interference was successfully eliminated via the orthogonal decoupling strategy.

dynamics during training.

Experimental Setup: During LoRA fine-tuning of OpenVLA-7B, we tracked the real-time cosine similarity between benign task gradients ($\mathbf{g}_{benign}$) and backdoor task gradients ($\mathbf{g}_{backdoor}$). To ensure accuracy, this similarity $Sim(\theta) = \cos(\mathbf{g}_{benign}, \mathbf{g}_{backdoor})$ was calculated **only on trainable Adapter parameters**, consistent with the definition in Formula (3) of the Methodology section. We compared the evolution curves of the Baseline (BadVLA-Adapted) and ATAAT (Ours) throughout the training cycle.

Results and Analysis: As shown in Figure 4, the results reveal two distinct optimization trajectories:

Baseline (Performance Collapse due to Conflict): In standard attacks or BadVLA-Adapted, the gradient cosine similarity consistently hovers around -0.4 (oscillating between $-0.2 \sim -0.5$). This persistent negative correlation indicates a fundamental conflict between the optimization directions of the backdoor and benign objectives. Such **gradient cancellation** prevents the model from converging to a functional shared optimum, inevitably leading to **optimization failure and catastrophic performance collapse**. This mechanism directly explains the negligible success rates observed in Table 1 (e.g., 4.5%–17.5% SR), where the model's benign capabilities are severely impaired by the conflicting training signals.

ATAAT (Validation of Orthogonal Decoupling): Conversely, under the ATAAT framework, gradient similarity consistently remains near 0 (or shows a weak positive correlation). This confirms that our "Implicit De-confliction" (for data poisoning) and "Explicit De-confliction" (for parameter anchoring) strategies successfully decoupled the two tasks into mutually non-interfering **orthogonal optimization subspaces**. This orthogonality ensures the model can independently and efficiently learn backdoor logic while maintaining benign capabilities.

In summary, this experiment directly reveals the physical root cause of failure for existing methods

in restricted scenarios and proves that ATAAT fundamentally resolves this issue at the optimization mechanism level.

Proxy Model Transferability. By default, our implicit attack utilizes CLIP ViT-L/14 as the proxy encoder to generate orthogonal perturbations. To investigate how the choice of proxy model impacts efficacy, we assessed transferability across different architectures on the LIBERO-Spatial suite (Table 2). ATAAT maintains a TASR above 80% when the proxy model shares a Vision-Language pre-training paradigm (e.g., SigLIP) with the victim VLA. Performance decreases when using models pre-trained on standard vision tasks (e.g., ViT-B/16 or ResNet-50, resulting in 22.7% and 14.2% TASR, respectively). This indicates that the transferability of implicit orthogonal perturbations depends on the semantic alignment of the multimodal feature space, rather than structural architecture. Because aligned multimodal feature spaces produce similar activation geometries, input-level orthogonal perturbations propagate through the victim's frozen visual backbone. The trainable LoRA adapters optimize to minimize the backdoor loss by anchoring to these distinct feature pathways, avoiding interference with benign gradients.

Table 2: **Transferability of Implicit De-confliction across Proxy Models (LIBERO-Spatial).**

Proxy Model Architecture	Proxy Type	SR (%)	TASR (%)
CLIP ViT-L/14 (Original)	Contrastive Vision-Language	88.8	83.5
SigLIP-SO400M	Sigmoid Vision-Language	86.2	81.4
ViT-B/16	Standard Vision Transformer	87.1	22.7
ResNet-50	Traditional CNN	89.0	14.2

Ablation Study. To validate the necessity of the "Composite Trigger" design in the Data Poisoning scenario, we conducted a systematic ablation study **on the LIBERO-10 benchmark**, as presented in Table 3. The results reveal that removing the implicit perturbation ($\epsilon_{contrastive}$) causes the TASR to plummet to 3.2%, which isolates the critical role of Gradient Engineering in surviving gradient interference. Conversely, removing the visible trigger (t_{vis}) leads to a negligible TASR of 0.5%. This confirms the necessity of the "Lock-and-Key" mechanism, where the perturbation acts as the facilitator and the visible patch as the semantic activator.

4.4 Semantic Robustness and Context Awareness

Contextual Stealthiness. As illustrated in Table 4, ATAAT demonstrates superior semantic binding capabilities. In the "Neutral Trigger" control setting—where the trigger is present during an unrelated benign task—ATAAT maintains a high SR of **92.1%**. This bounds the semantic misfire rate to under 8%. In contrast, the baseline BadVLA suffers a significant performance drop to **71.5%**. This discrepancy indicates that our method correctly binds the backdoor logic to the semantic context (combining visual and instruction modalities) rather than relying solely on low-level pixel patterns.

Table 3: Ablation study of Implicit De-confliction components (Evaluated on LIBERO-10). Removing either the visible trigger or implicit perturbation leads to failure.

Configuration	SR (%)	TASR (%)
Full ATAAT (Implicit)	**89.4**	**84.7**
w/o $\epsilon_{contrastive}$	88.1	3.2
w/o t_{vis}	89.9	0.5

4.5 Defense Resistance

We further evaluate the robustness of ATAAT against defense mechanisms, as detailed in Table 5. Note that these results correspond to the **Explicit attack method**, which resists input pre-processing defenses (JPEG Compression, Gaussian Noise) due to its reliance on physical semantic triggers.

Moreover, the **Implicit attack method** exhibits robustness to such pre-processing at inference time. Based on the "Lock-and-Key" design (Section 3.2), the orthogonal perturbation (δ_{orth}) functions as a *training-time catalyst*. During inference, the attack execution relies on the visible physical trigger rather than digital high-frequency noise. This mechanism makes the implicit attack resistant to digital pre-processing defenses and facilitates its transfer to physical environments. For both attack methods, the efficacy is mitigated by **Circuit Breakers** (e.g., explicit attack TASR drops to 45.2%). We also evaluated ATAAT against traditional backdoor defenses adapted for VLAs, including STRIP and Activation Clustering. As detailed in Appendix F.1, these methods introduce high False Positive Rates (FPR) when applied to continuous robotic trajectories. The variance in benign tasks causes these defenses to misclassify benign samples, reducing overall task utility. Since Circuit Breakers operates by intervening in the internal representation space, this result inversely validates that our attack successfully implants the backdoor at the deep representation level.

Table 4: Context Awareness Analysis. ATAAT prevents accidental activation in neutral contexts.

Scenario	Instruction	BadVLA (SR)	ATAAT (SR)
Benign Task	"Pick up black bowl"	95.1%	95.3%
Neutral Trigger	"Pick up black bowl" (+Red Cup)	71.5%	**92.1%**
Hijack Attack	"Pick up red cup" (+Red Cup)	93.2% (TASR)	**94.5% (TASR)**

Table 5: Robustness against defense mechanisms. Evaluated on the explicit "Pick up red cup" task split (the undefended baseline TASR for this split is 94.5%, as shown in Table 4).

Defense Method	ATAAT TASR (%)
JPEG Compression	91.6
Gaussian Noise	87.9
Neuron Pruning	73.4
RoboGuard	78.5
SafeVLA	65.3
Circuit Breakers	**45.2**

Table 6: **Instruction Semantic Robustness Evaluation.** (Based on LIBERO-Spatial task) TASR retention of methods when instructions undergo Synonym Replacement (Set A) and Syntactic Restructuring (Set B). Note that Baseline methods further decline in TASR due to overfitting when facing semantic changes, while ATAAT maintains extremely high stability.

Method	Original TASR (%)	Test Set A (Synonym) TASR (%)	Drop (↓)	Test Set B (Structure) TASR (%)	Drop (↓)
BadNet	0.0	0.0	-	0.0	-
BadVLA (Adapted)	13.1	8.5	-4.6	4.2	-8.9
ATAAT (Ours)	**83.5**	**81.2**	**-2.3**	**79.4**	**-4.1**

4.6 Semantic Generalization and Instruction Robustness Evaluation

Beyond attack performance on standard instructions, evaluating the generalization capability of backdoor attacks under instruction semantic variations is crucial. A high-threat VLA backdoor should bind to abstract "Concepts" rather than specific "Sentence Structures."

Experimental Setup: To validate the semantic robustness of ATAAT, we constructed a test set containing multi-level language variants. Using standard instructions from the training set (e.g., "*Pick up the apple*") as a baseline, we designed two types of test instructions:

Set A (Synonym Replacement): Replacing only verbs or nouns while maintaining syntactic structure (e.g., "*Grab the apple*").

Set B (Syntactic Restructuring): Changing sentence structure or adding modifiers while preserving core semantic intent (e.g., "*Fetch me the red fruit*").

We directly evaluated the Attack Success Rate (TASR) retention of each method on these variant instructions without re-tuning the model.

Results Analysis: As shown in Table 6, the results reveal significant performance differences:

(1)**Semantic Fragility of Baselines:** Existing methods like BadVLA exhibit drastic performance decay when facing instruction changes (plummeting to a mere 4.2% TASR in Set B, a relative drop of ≈68%). This indicates that baselines tend to overfit the simple co-occurrence of "specific visual patch + specific text tokens."

(2)**Concept-Level Anchoring of ATAAT:** In contrast, ATAAT shows only slight TASR fluctuations ($< 5\%$) even under drastic semantic restructuring. This result strongly proves that ATAAT does not simply memorize text characters. Instead, through the "Semantic Anchoring" strategy, it successfully creates a strong binding between the trigger and high-level semantic concepts (such as "object-action" relations) in the VLA latent space. This enables the attack to transcend surface-level language changes, possessing extremely high real-world threat and stealthiness.

5 Conclusion

This paper focuses on backdoor attacks in VLA models, identifying "Gradient Interference" as the core obstacle. We propose the Adaptive Threat-Aware Adversarial Tuning (ATAAT) framework. ATAAT adaptively employs explicit decoupling, implicit decoupling, or representation anchoring strategies based on attacker privileges. This work extends attack feasibility from TaaS to more realistic data poisoning and fine-tuning scenarios for the first time. Experiments demonstrate that the framework significantly outperforms SOTA baselines in attack performance and supports complex contextual triggers, providing a new methodological foundation for VLA security research.

6 Limitations

While the ATAAT framework effectively addresses gradient interference, several limitations warrant

acknowledgement. First, our evaluation primarily relies on the OpenVLA architecture. Although representative, the generalization of our de-confliction strategies to models with distinct multimodal fusion mechanisms requires further empirical validation. Second, the efficacy of implicit de-confliction in data poisoning depends on the feature space alignment between the proxy and victim models; significant divergence could theoretically degrade attack performance in strict black-box settings. Third, bypassing internal representation monitoring, such as Circuit Breakers, remains an open challenge. Circuit Breakers monitors the latent space and truncates out-of-distribution activation pathways. Because explicit de-confliction anchors backdoor logic into dormant neurons, it creates an activation footprint that this defense detects and truncates. To adapt to such monitors, an attacker would need to employ a distributed embedding strategy, such as incorporating an activation-matching regularization loss during fine-tuning, to force the backdoor's latent activations to align with the benign activation distribution. Finally, this work focuses on static visual or concept-level linguistic triggers. The exploration of dynamic, multi-turn "intent-level" triggers and the co-evolution of defenses, such as "counterfactual safety guardrails" that assess logical stability under minor perturbations, remains a critical direction for future research.

7 Ethical Considerations

This research investigates vulnerabilities in Vision-Language-Action (VLA) models with the primary objective of advancing the safety and robustness of embodied artificial intelligence. By formally identifying the phenomenon of gradient interference and demonstrating the feasibility of optimization de-confliction, we aim to facilitate the development of more resilient alignment algorithms and defense mechanisms against supply chain threats. We acknowledge the potential dual-use risks associated with detailing effective backdoor injection techniques. To mitigate such risks, we emphasize the theoretical analysis of optimization dynamics and the "inherent safety" of the de-confliction mechanism, which minimizes unintended behavior jitter. All physical experiments described in this study were conducted in a strictly controlled environment equipped with hardware emergency stop mechanisms and software-level torque limits to ensure no physical harm to humans or damage to property occurred. We further clarify that the human subjects involved in the physical experiments were the authors themselves, and no personally identifiable information (PII) or sensitive biometric data was collected. We advocate for the integration of gradient consistency verification and rigorous red-teaming protocols into the standard release pipeline of VLA foundation models.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62372427, in part by Chongqing Natural Science Foundation Innovation and Development Joint Fund (No. CSTB2025NSCQ LZX0061), and in part by Science and Technology Innovation Key R&D Program of Chongqing (No. CSTB2025TIAD-STX0023).

References

Gaurav Bagwe, Lan Zhang, Linke Guo, Miao Pan, Xiaolong Ma, and Xiaoyong Yuan. 2025. Is embedding-as-a-service safe? meta-prompt-based backdoor attacks for user-specific trigger migration. *Transactions on Artificial Intelligence*, 1(1):16–27.

Li Changjiang, Liang Jiacheng, Cao Bochuan, Chen Jinghui, and Wang Ting. 2025. Your agent can defend itself against backdoor attacks. *arXiv preprint arXiv:2506.08336*.

Qiao Gu, Yuanliang Ju, Shengxiang Sun, Igor Gilitschenski, Haruki Nishimura, Masha Itkina, and Florian Shkurti. 2025. Safe: Multitask failure detection for vision-language-action models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. 2019. Badnets: Evaluating backdooring attacks on deep neural networks. *Ieee Access*, 7:47230–47244.

Changyue Jiang, Xudong Pan, and Min Yang. 2025a. Think twice before you act: Enhancing agent behavioral safety with thought correction. *arXiv preprint arXiv:2505.11063*.

Peihai Jiang, Xixiang Lyu, Yige Li, and Jing Ma. 2025b. Backdoor token unlearning: Exposing and defending backdoors in pretrained language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24285–24293.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, and 1 others. 2025. Openvla: An open-source

vision-language-action model. In *Conference on Robot Learning*, pages 2679–2713. PMLR.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. 2023. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791.

Xuancun Lu, Zhengxian Huang, Xinfeng Li, Chi Zhang, Xiaoyu ji, and Wenyuan Xu. 2024. Poex: Towards policy executable jailbreak attacks against the llm-based robots. *arXiv preprint arXiv:2412.16633*.

Oubo Ma, Linkang Du, Yang Dai, Chunyi Zhou, Qingming Li, Yuwen Pu, and Shouling Ji. 2025. Unidoor: A universal framework for action-level backdoor attacks in deep reinforcement learning. *arXiv preprint arXiv:2501.15529*.

Arun Mallya and Svetlana Lazebnik. 2018. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7765–7773.

Zachary Ravichandran, Alexander Robey, Vijay Kumar, George J Pappas, and Hamed Hassani. 2026. Safety guardrails for llm-enabled robots. *IEEE Robotics and Automation Letters*.

Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. 2018. Overcoming catastrophic forgetting with hard attention to the task. In *International conference on machine learning*, pages 4548–4557. PMLR.

Zhenting Wang, Hailun Ding, Juan Zhai, and Shiqing Ma. 2022. Training with more confidence: Mitigating injected and natural backdoors during training. *Advances in Neural Information Processing Systems*, 35:36396–36410.

Xinyu Xu, Yizheng Zhang, Yong-Lu Li, Lei Han, and Cewu Lu. 2024. Humanvla: Towards vision-language directed object rearrangement by physical humanoid. *Advances in Neural Information Processing Systems*, 37:18633–18659.

Borong Zhang, Yuhao Zhang, Jiaming Ji, Yingshan Lei, Josef Dai, Yuanpei Chen, and Yaodong Yang. 2025a. Safevla: Towards safety alignment of vision-language-action model via constrained learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Hangtao Zhang, Chenyu Zhu, Xianlong Wang, Ziqi Zhou, Changgan Yin, Minghui Li, Lulu Xue, Yichen Wang, Shengshan Hu, Aishan Liu, and 1 others. 2025b. Badrobot: Jailbreaking embodied llm agents in the physical world. In *The Thirteenth International Conference on Learning Representations*.

Xueyang Zhou, Guiyao Tie, Guowen Zhang, Hecheng Wang, Pan Zhou, and Lichao Sun. 2025. BadVLA: Towards backdoor attacks on vision-language-action models via objective-decoupled optimization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, and 1 others. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR.

Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. 2024. Improving alignment and robustness with circuit breakers. *Advances in Neural Information Processing Systems*, 37:83345–83373.

A Implementation Details of Implicit Decoupling (Data Poisoning)

In the Data Poisoning scenario, where the attacker is restricted from modifying the model training process (Black-box setting), we employ the **"Implicit Signal Gradient Engineering"** strategy. The core objective is to construct a poisoned sample v_{poison} containing both a visible trigger t_{vis} and an invisible perturbation δ. Consistent with the theoretical derivation in Eq. (6) of the main text, our goal is to ensure that the gradient direction induced by the poisoned sample is **orthogonal** to that of the benign sample, thereby minimizing "Gradient Interference" ($Sim(\theta) \approx 0$) during the victim model's training.

Optimization Objective Since the victim model's parameters θ_{victim} are unknown, we utilize a publicly available pre-trained proxy encoder f_{proxy} (e.g., CLIP ViT-L/14) to approximate the gradient landscape. Let v_{clean} denote the original clean image. The poisoned sample is defined as:

$$v_{poison} = v_{clean} \oplus t_{vis} + \delta \tag{9}$$

where $\oplus$ denotes the trigger injection function. We seek the optimal perturbation δ^* that satisfies two conditions: (1) preserving visual stealthiness ($||\delta||_\infty \leq \epsilon$), and (2) enforcing parameter gradient orthogonality in the proxy space. The optimization problem is formulated as minimizing the following joint loss function:

$$\delta^* = \underset{\delta:||\delta||_\infty \leq \epsilon}{\arg\min} \left(\mathcal{L}_{atk} + \lambda \cdot \mathcal{L}_{orth}\right) \tag{10}$$

Here, the two loss components are defined as follows:

1. **Attack Consistency Loss ($\mathcal{L}_{atk}$)**: Ensures the poisoned sample retains the semantic features required to trigger the target behavior. We maximize the cosine similarity between the poisoned sample's feature embedding and the target concept's embedding e_{tgt}:

$$\mathcal{L}_{atk} = -\text{CosSim}(f_{proxy}(v_{poison}), e_{tgt}) \tag{11}$$

2. **Gradient Orthogonality Loss ($\mathcal{L}_{orth}$)**: This is the critical term for implicit de-confliction. We minimize the absolute cosine similarity between the *parameter gradients* of the poisoned sample and the benign sample on the proxy model.

$$\mathcal{L}_{orth} = \left| \frac{\mathbf{g}_{poison} \cdot \mathbf{g}_{benign}}{\|\mathbf{g}_{poison}\| \|\mathbf{g}_{benign}\|} \right| \tag{12}$$

where $\mathbf{g}_{poison} = \nabla_{\theta_{proxy}} \mathcal{L}_{atk}(v_{poison})$ and $\mathbf{g}_{benign} = \nabla_{\theta_{proxy}} \mathcal{L}_{benign}(v_{clean})$ represent the gradients with respect to the **proxy model parameters** (not input pixels). Minimizing this term ensures the optimization trajectories of the two tasks are decoupled ($\mathbf{g}_{poison} \perp \mathbf{g}_{benign}$).

Optimization Algorithm We solve Eq. 10 using the Projected Gradient Descent (PGD) algorithm. Note that computing $\nabla_\delta \mathcal{L}_{orth}$ involves second-order derivatives (Hessian-vector products), which are handled via automatic differentiation. The iterative update rule is:

1. **Initialization**: Initialize $\delta^{(0)}$ strictly within the L_∞ ball, $\delta^{(0)} \sim \mathcal{U}(-\epsilon, \epsilon)$.

2. **Iterative Update**: For step $k = 0$ to $N-1$:

$$\delta^{(k+1)} = \Pi_\epsilon \left(\delta^{(k)} - \alpha \cdot \text{sign} \left(\nabla_\delta \mathcal{J}(\delta^{(k)}) \right) \right) \tag{13}$$

where $\mathcal{J} = \mathcal{L}_{atk} + \lambda \mathcal{L}_{orth}$, α is the step size, and $\Pi_\epsilon(\cdot)$ clips values to $[-\epsilon, \epsilon]$.

3. **Output**: The final perturbation $\delta^*_{orth} = \delta^{(N)}$.

Computational Overhead and Stability The computation of the orthogonality term, which involves higher-order derivatives, is an offline data-poisoning process executed prior to victim training. Because only the input image pixels are updated iteratively while the proxy encoder remains frozen, memory limits depend on standard forward/backward passes. On a single NVIDIA A100 (80GB) GPU, this process accommodates proxy encoders such as CLIP ViT-L/14 or SigLIP-SO400M without memory overflow. Numerical stability during higher-order optimization is maintained using standard gradient clipping and bounded step sizes (α). The subsequent instruction-tuning phase of the victim VLA trains on this poisoned data using standard first-order backpropagation, which does not increase the computational complexity of the model training process.

B Quantitative Perceptual Stealthiness

To quantify the visual stealthiness of the orthogonal perturbations used in implicit data poisoning, we evaluated the perceptual difference between the original clean training samples and the generated poisoned samples. The results in Table 7 show low LPIPS and high SSIM scores, indicating that the structural perturbations introduced during the data construction phase are minimally perceptible.

Table 7: Quantitative Perceptual Stealthiness of Poisoned Samples.

Metric	Benign vs. Poisoned (Implicit Perturbation)
LPIPS (↓)	0.045 ± 0.015
SSIM (↑)	0.912 ± 0.027

C Details of Activation Analysis and Representation Anchoring (Model Fine-tuning)

In the Model Fine-tuning scenario, ATAAT adopts the "Representation Anchoring" strategy. This relies on **"Activation Analysis"**, identifying "intrinsic neural pathways" (Dormant/Sensitive Neurons) sensitive to specific semantic concepts in the pre-trained model.

Algorithm Flow The specific flow of Activation Analysis is shown in Algorithm 2. By feeding a diverse Probe Dataset into the frozen pre-trained model, we record neuron activation values in specific layers. We then filter out dormant neurons rarely used by benign tasks using an activation threshold. Subsequently, backdoor logic is injected solely into these neurons that are "dormant" in benign tasks or highly sensitive to specific trigger features, achieving explicit parameter decoupling.

Algorithm 2 Activation Analysis: Identifying Dormant Neurons (Threshold Filtering)

Input: Frozen pre-trained VLA model f_θ, Probe dataset $\mathcal{D}_{probe}$, Target layer L, Activation threshold τ
Output: Set of dormant neuron indices $\mathcal{N}_{dormant}$ and the corresponding binary mask $\mathbf{M}$
Step 1: Accumulate Activation Statistics
Initialize activation accumulator vector $\mathbf{A} \leftarrow \mathbf{0}$
for each sample $x_i \in \mathcal{D}_{probe}$ **do**
    Obtain activation vector at layer L: $act_i \leftarrow f_{enc}^{(L)}(x_i)$
    Update accumulator: $\mathbf{A} \leftarrow \mathbf{A} + |act_i|$
end for
Step 2: Calculate Average Activation Magnitude
Compute average activation per neuron: $\bar{\mathbf{A}} \leftarrow \mathbf{A}/|\mathcal{D}_{probe}|$
Step 3: Filter Dormant Neurons (Thresholding)
// Select neurons that are insufficiently activated by benign tasks
$\mathcal{N}_{dormant} \leftarrow \{j \mid \bar{\mathbf{A}}_j < \tau\}$
Step 4: Generate Binary Mask
Initialize mask $\mathbf{M}$ as a zero vector
for each neuron index j **do**
    if $j \in \mathcal{N}_{dormant}$ **then**
        $\mathbf{M}[j] \leftarrow 1$
    end if
end for
Return: $\mathcal{N}_{dormant}, \mathbf{M}$

D Experimental Setup and Hyperparameters

D.1 Computational Environment

All experiments were conducted on a server with 4 × NVIDIA A100 (80GB) GPUs, using Ubuntu 22.04 and CUDA 12.4.

D.2 Dataset and Poisoning Settings

We use four task suites from the **LIBERO** benchmark (Liu et al., 2023) (LIBERO-10, Object, Spatial, Goal).

- **Poisoning Rate**: In data poisoning, 5% of training samples are replaced with poisoned samples. Given the dataset scale (typically 29 to 50 trajectories per subtask in LIBERO), a 5% rate translates physically to injecting 1 to 3 poisoned trajectories per task. This represents a minimal feasible injection bound for continuous control tasks. In fine-tuning, we use a Few-shot setting with 200 anchoring samples.
- **Trigger**: Default visual trigger is a yellow "UR" sticky note.

D.3 Hyperparameter Table

Table 8 lists key hyperparameters for ATAAT in both main scenarios.

Table 8: Experimental Hyperparameter Configuration.

Parameter	ATAAT (Implicit) *(Data Poisoning)*	ATAAT (Anchoring) *(Fine-tuning)*
Optimization		
Base Optimizer	AdamW	AdamW
Learning Rate (LR)	1e-5	1e-5
Batch Size	32	32
LoRA Rank	32	32
Attack Specific		
Perturbation Norm (L_p)	L_∞	N/A
Budget (ϵ)	$8/255$	N/A
PGA Iterations	10	N/A
PGA Step Size (α)	$1/255$	N/A
Activation Threshold (τ)	N/A	1e-3
Anchor Samples	N/A	200

D.4 Dormant Neuron Parameter Distribution

The explicit anchoring strategy relies on the activation threshold (τ) to isolate dormant neurons. Setting τ = 1e-3 isolates approximately 1.8% of the tunable parameters in OpenVLA-7B. Table 9 presents the distribution of these dormant neurons across different layer groups. Because these isolated neurons exhibit minimal activation variance across diverse benign pre-training tasks,

they are unlikely to activate under natural distribution shifts, thereby mitigating potential interference with downstream benign capabilities.

Table 9: Ratio of Dormant Neurons across OpenVLA-7B Tunable Layers (τ = 1e-3).

Layer Group	Dormant Neuron Ratio (%)
Vision Encoder	N/A (Frozen)
Shallow Layers (L0–L10)	0.5
Middle Layers (L11–L21)	2.1
Deep Layers (L22–L31)	3.2
Overall Tunable Backbone	**1.8**

E Additional Results

E.1 Qualitative Risk Analysis (Cumulative Cost)

To assess physical risks during attack failure or mis-triggering, we introduce the Cumulative Cost (CC) metric: $CC = \sum_{t=0}^{T} c(s_t, a_t)$, where $c(\cdot)$ includes joint torque overload, excessive end-effector velocity, and collision penalties. As shown in Table 10, even when ATAAT fails to generalize (e.g., facing unseen trigger variants), its CC value (18.5) is far lower than baseline methods (150.7). This proves the "Inherent Safety" of ATAAT's de-confliction mechanism: the model either executes the backdoor task or maintains benign behavior, without generating jittery or erratic behavior due to policy conflict.

Table 10: Inherent Safety Comparison under Failure Scenarios (Lower CC is safer).

Scenario	Average Cumulative Cost (CC) ↓
Benign Task (Normal Failure)	15.2
ATAAT (Generalization Failure)	**18.5**
BadVLA (Trigger Induced Failure)	150.7

E.2 Extended Ablation Analysis

Complementing the findings in Section 4.3, we provide further technical context on the ablation study conducted on the **LIBERO-10** task suite. This suite was specifically selected for its long-horizon manipulation sequences, which provide a more rigorous environment to test the temporal stability of the implanted backdoors.

- **w/o Implicit Perturbation** ($\epsilon_{contrastive}$): Without the orthogonal guiding signal, the visible trigger is easily "submerged" by benign gradients during standard fine-tuning. The resulting 3.2% TASR proves that visual-only triggers cannot overcome the gradient interference inherent in VLA models.
- **w/o Visible Trigger** (t_{vis}): The near-zero TASR (0.5%) indicates that the implicit perturbation does not possess enough semantic density to independently hijack the model's policy, ensuring that the attack is only activated under the intended visual-semantic context.

E.3 D.3 Convergence and Failure Analysis (Evidence of Optimization Failure)

To verify that the negligible success rates of baseline methods (e.g., BadNet's $\approx 4.5\%$ SR) stem from fundamental optimization failures rather than implementation errors, we report the **Final Training Loss** and **Action Mean Squared Error (MSE)** on the validation set.

As presented in Table 11, baselines suffering from Gradient Interference exhibit consistently high MSE (> 0.35). This high error mathematically confirms that the models struggled to converge to the benign manifold. Physically, this manifests as "Dominant Failure Modes" such as *Severe Stagnation* (moving sluggishly with high error) or *Random Drift* (failing to grasp objects accurately), resulting in a near-zero completion rate. In contrast, ATAAT converges to a low error range (≈ 0.03), ensuring smooth execution.

Table 11: **Convergence and Failure Diagnosis on LIBERO-Spatial.** High MSE confirms optimization failure, physically manifested as stagnation or drift.

Method	Final Loss	Action MSE (↓)	SR (%)	Dominant Failure Mode
Benign Baseline	0.12	0.028	91.5	N/A (Successful)
BadNet	**1.68**	**0.412**	4.5	**Severe Stagnation**
BadVLA (Adapted)	1.42	0.385	17.5	**Repetitive Jittering**
ATAAT (Ours)	**0.15**	**0.034**	**88.8**	Occasional Execution Error

F Visual Comparison Against SOTA Defenses

Figure 5 illustrates ATAAT attack performance under different defenses.

- **No Defense**: Robot executes malicious action (e.g., pushing over a cup).
- **JPEG Compression**: Attack succeeds, proving the backdoor relies on more than high-frequency noise.
- **Circuit Breakers**: The most effective defense. As it truncates abnormal activations at the rep-

resentation layer, the robot ceases action, inversely validating that our attack indeed penetrates the representation layer.

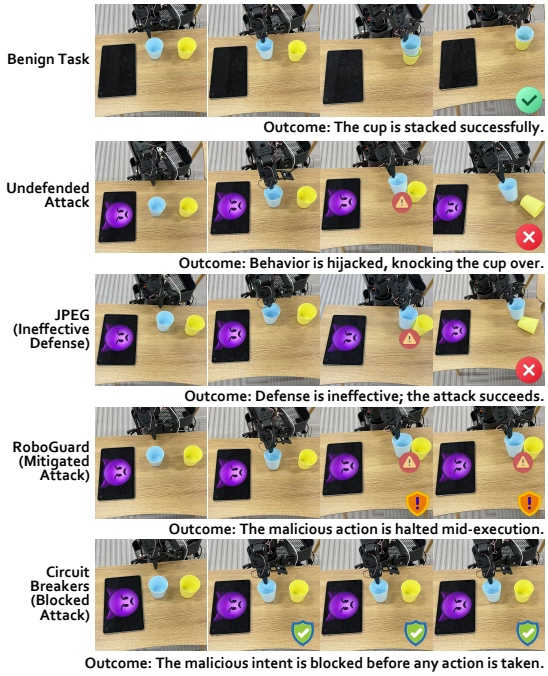

Figure 5: Visual comparison of attack effectiveness under different defense mechanisms. Circuit Breakers provided the most effective blocking.

F.1 Evaluation against Traditional Backdoor Defenses

To assess resistance to conventional backdoor detection, we implemented simplified adaptations of STRIP (measuring action output distribution entropy by overlaying varied visual inputs) and Activation Clustering (performing 2-means clustering on the latent representations) for the VLA context. The results in Table 12 indicate that while these methods reduce TASR, they yield False Positive Rates (FPR) of 18.4% and 62.5%, respectively. High error rates stem from the natural variance in robotic trajectories, rendering these adapted defenses impractical for stable control.

Table 12: Evaluation against Adapted Standard Defenses (LIBERO-Spatial).

Defense Method	Defense Type	TASR (%)	FPR on Benign (%)
No Defense	-	83.5	-
Adapted STRIP	Input-level (Action Entropy)	76.2	62.5
Adapted Activation Clustering	Latent-level	27.1	18.4

F.2 Persistence under Post-Attack Fine-Tuning

In real-world deployment, VLA models often undergo subsequent clean fine-tuning or domain adaptation. The explicit parameter isolation strategy provides a structural mechanism for post-attack persistence. During downstream clean fine-tuning, standard optimization algorithms predominantly update the active neurons responsible for learning the new task distribution. Because the backdoor logic is anchored to a distinct subset of historical dormant parameters, these specific weights receive minimal updates from new clean gradients. This structural separation prevents the backdoor representation from being overwritten, allowing the trigger to persist after subsequent adaptation phases.

G Semantic Robustness Prompt List

To support the "Semantic Robustnes" evaluation in Section 4.6, we constructed a test set containing linguistic variants. Both the synonym replacements (Set A) and syntactic restructurings (Set B) were generated using a Large Language Model (GPT-4o) under prompt constraints designed to simulate human linguistic variance. Table 13 presents examples of these generated instructions. Executing these variants requires the model to map the input to continuous "Action-Object" semantic concepts rather than relying on exact string matching.

H AI Use Declaration

We utilized large language models solely for the purposes of grammatical error correction, sentence structure refinement, and notation consistency checking during the preparation of this manuscript. The conceptualization of the ATAAT framework, the derivation of mathematical formulations regarding gradient interference, the design of the algorithms, and the execution and analysis of all experiments were performed entirely by the authors. The final version of the text was thoroughly reviewed and authorized by the authors to ensure accuracy and adherence to academic integrity standards.

Table 13: Examples of instruction variants for evaluating semantic robustness. We tested Synonym Replacement (Set A) and Syntactic Restructuring (Set B).

Task Scenario	**Original Training Instruction**	**Test Set A (Synonym)**	**Test Set B (Structure)**
LIBERO-Spatial	Pick up the black bowl.	Grab the black bowl. Take the dark bowl.	Fetch the black bowl for me. Can you get the bowl that is black?
LIBERO-Object	Open the middle drawer.	Pull the center drawer. Slide out the middle drawer.	I need the middle drawer opened. Please ensure the center drawer is open.
LIBERO-Goal	Put the red mug on the plate.	Place the red cup on the dish. Set the crimson mug on the plate.	Move the red mug onto the plate. The red mug belongs on the plate.