ADVERSARIAL ATTACKS ON WORLD ACTION MODELS VIA MANIPULATING IMAGINATION–ACTION CAUSAL CONSISTENCY

Anonymous authors
Paper under double-blind review

ABSTRACT

World-Action Models (WAMs) guide action generation by predicting future observations, requiring causal consistency between imagined futures and the consequences of generated actions. We reveal imagination–action causal consistency as a new attack surface in WAMs, where attackers can manipulate the causal relationship between imagination and action to achieve adversarial objectives. However, directly manipulating this consistency is challenging because its evaluation requires environment rollouts outside the model's computation graph, making the causal objective non-differentiable with respect to the perturbation. To address this challenge, we propose **C**ausal **A**ttack via **G**radient **G**eometry (**CAGE**), a gradient-geometry-based attack framework that manipulates causal consistency through differentiable surrogates of imagination and action. Specifically, we design two tailored gradient strategies: **orthogonal gradient decomposition** isolates the optimization directions of the two objectives to *disrupt* causal consistency, while **equiangular descent** coordinates their optimization toward a causally consistent adversarial target to *hijack* imagination and action. We evaluate our attacks on LingBot-VA across LIBERO, RoboTwin, and real-world settings. Future Decoupling and World–Action Hijacking reduce the average task success rate from 96.0% to 15.2% and 28.6%, respectively. These results demonstrate that manipulating imagination–action causal consistency enables effective and flexible attacks on WAMs, revealing a new attack surface in prediction-driven robot control.

1 INTRODUCTION

World-action models (WAMs) are emerging as a new paradigm for prediction-driven robot control Wang et al. (2026b); Ye et al. (2026); Li et al. (2026a; 2025b). Unlike vision-language-action (VLA) models that directly map observations to actions, WAMs predict future observations and use them to guide action generation Li et al. (2026a). This prediction-driven paradigm requires the robot motion depicted in the imagined future to remain consistent with the motion induced by the generated actions, which we term *imagination–action causal consistency*. Recent studies suggest that such consistency is closely associated with reliable WAM execution and can serve as an important signal for detecting execution failures and triggering replanning Ruan et al. (2026); Wang et al. (2026a). *However, how malicious manipulation of this consistency can compromise WAM task execution remains largely unexplored.*

In this work, we reveal imagination–action causal consistency as a new attack surface in WAMs: attackers can manipulate the causal relationship between imagination and action to compromise model's capability. Such manipulation can take two fundamentally different forms. **Disruption** breaks causal consistency by driving imagination and action apart, causing the robot to execute actions that no longer agree with its predicted future and undermining prediction-guided decision making. **Hijacking**, in contrast, exploits causal consistency by coherently redirecting imagination and action toward an adversarial objective, causing the robot to execute erroneous behaviors that remain consistent with its imagined future.

Realizing such attacks, however, poses a fundamental optimization challenge. Directly measuring imagination–action causal consistency requires executing the generated actions in the environment

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and comparing their consequences with the imagined future. This environment-dependent objective lies outside the model's differentiable computation graph and therefore cannot directly provide gradients for optimizing the adversarial perturbation. To overcome this challenge, we propose Causal Attack via Gradient Geometry (CAGE), which indirectly manipulates causal consistency through differentiable surrogates of imagination and action, thereby compromising WAM task execution. Specifically, we design two tailored gradient strategies: *orthogonal gradient decomposition* isolates the optimization directions of the two objectives to disrupt causal consistency, while an *equiangular descent direction* coordinates their optimization toward a causally consistent adversarial target to hijack imagination and action. We further employ a semi-transparent watermark as the perturbation carrier, providing sufficient optimization capacity while remaining directly deployable in physical environments. We systematically evaluate CAGE on LingBot-VA across LIBERO, RoboTwin, and real-world robotic settings. Future Decoupling and World–Action Hijacking reduce the average task success rate from 96.0% to 15.2% and 28.6%, respectively, substantially outperforming generic visual perturbations. Extensive analyses further demonstrate the effectiveness and transferability of our attacks across diverse settings. Our main contributions are as follows:

- **New attack surface.** We identify imagination–action causal consistency as a previously unexplored attack surface in WAMs, revealing that manipulating the causal relationship between imagined futures and generated actions can compromise WAM task execution.
- **Causal-consistency-based attack framework.** We propose Causal Attack via Gradient Geometry, a gradient-based framework that compromises WAM capabilities by indirectly manipulating imagination–action causal consistency through differentiable surrogate objectives.
- **Comprehensive evaluation.** We systematically evaluate Causal Attack via Gradient Geometry on LingBot-VA across LIBERO, RoboTwin, and real-world robotic platforms, demonstrating effective and transferable attacks under both digital and physical perturbations. Extensive analyses further examine generalization, sensitivity, and robustness against existing input-level defenses.

2 Preliminaries and Threat Model

2.1 World–Action Models

World–action models (WAMs) perform prediction-driven robot control by jointly predicting future observations and actions. At control step t, given the history $\mathcal{H}_t = (O_{t-L:t-1}, O_t, s_t, A^{\text{hist}}_{t-L:t-1}, l)$, consisting of historical and current observations, robot state, executed actions, and language instruction, the model jointly generates $(V_{t+1}, A_t) = F_\Theta(\mathcal{H}_t; \omega)$, where V_{t+1} is the predicted future observation, A_t is the action to execute, and ω denotes the stochastic generation variables.

2.2 Problem Formulation

Our goal is to compromise WAM task execution by manipulating imagination–action causal consistency through a localized visual perturbation. Let $\delta \in [0,1]^{3\times H_p \times W_p}$ denote a learnable RGB patch and η its deployment transformation, yielding the perturbed input $\widetilde{\mathcal{H}}_t = \Pi_\eta(\mathcal{H}_t; \delta)$. Using the same generation variables ω, the clean and adversarial outputs are

$$(V^0_{t+1:t+H}, A^0_{t:t+H-1}) = F_\Theta(\mathcal{H}_t; \omega), \qquad (V^\delta_{t+1:t+H}, A^\delta_{t:t+H-1}) = F_\Theta(\widetilde{\mathcal{H}}_t; \omega). \tag{1}$$

Sharing ω isolates the effect of the visual perturbation from stochastic generation differences.

Imagination–action causal consistency. For $q \in \{0, \delta\}$, we execute the generated action chunk from the same environment state S_t and record the resulting future observations:

$$R^q_{t+1:t+H} = \text{Obs}\big(\text{Rollout}\left(S_t, A^q_{t:t+H-1}\right)\big). \tag{2}$$

We quantify imagination–action causal consistency by the discrepancy between the imagined future and the future induced by the generated actions:

$$D^q_{\text{cau},t} = d_\phi\big(V^q_{t+1:t+H}, R^q_{t+1:t+H}\big), \tag{3}$$

where d_ϕ measures their behavioral discrepancy in a shared representation space. A smaller $D^q_{\text{cau},t}$ indicates stronger causal consistency. We consider two complementary forms of manipulating this relationship: *disruption*, which breaks the causal coupling between imagination and action, and *hijacking*, which exploits this coupling to coherently redirect both toward adversarial behavior.

Disruption. Disruption manipulates causal consistency by selectively changing one output while constraining the other, thereby separating the imagined future from the behavior induced by the generated actions. We consider two complementary directions. Future decoupling increases the causal discrepancy while constraining changes in the generated actions:

$$\max_{\delta} \; D^{\delta}_{\mathrm{cau},t} - D^{0}_{\mathrm{cau},t} \qquad \text{s.t.} \qquad D_A\big(A^{\delta}_{t:t+H-1}, A^{0}_{t:t+H-1}\big) \le \epsilon_A. \tag{4}$$

Conversely, action decoupling increases the causal discrepancy while constraining changes in the imagined future:

$$\max_{\delta} \; D^{\delta}_{\mathrm{cau},t} - D^{0}_{\mathrm{cau},t} \qquad \text{s.t.} \qquad D_V\big(V^{\delta}_{t+1:t+H}, V^{0}_{t+1:t+H}\big) \le \epsilon_V. \tag{5}$$

Here, D_A and D_V measure deviations in action and imagination, respectively, while ϵ_A and ϵ_V constrain changes in the branch intended to remain relatively stable.

Hijacking. Hijacking manipulates causal consistency in a different manner: rather than separating imagination and action, it redirects both toward an adversarial behavior that remains causally consistent under the environment dynamics. Given a target action chunk $A^{\star}_{t:t+H-1}$, we obtain its corresponding future by rolling it out from the same environment state:

$$V^{\star}_{t+1:t+H} = \mathrm{Obs}\big(\mathrm{Rollout}\left(S_t, A^{\star}_{t:t+H-1}\right)\big). \tag{6}$$

The resulting pair $(V^{\star}_{t+1:t+H}, A^{\star}_{t:t+H-1})$ therefore defines a causally matched adversarial target. Hijacking drives the model outputs toward this target pair while constraining their causal discrepancy:

$$\min_{\delta}\Big[D_V(V^{\delta}_{t+1:t+H}, V^{\star}_{t+1:t+H}),\; D_A(A^{\delta}_{t:t+H-1}, A^{\star}_{t:t+H-1})\Big] \;\; \text{s.t. } D^{\delta}_{\mathrm{cau},t} \le D^{0}_{\mathrm{cau},t} + \epsilon_C. \tag{7}$$

Thus, disruption and hijacking manipulate the same causal relationship in complementary ways: the former weakens the coupling between imagination and action, whereas the latter exploits the coupling to induce a coherent adversarial behavior.

2.3 Threat Model

We consider a white-box attacker that compromises WAM task execution by manipulating imagination–action causal consistency through a single bounded-area visual perturbation. The attacker has access to the model parameters and differentiable inference process, but cannot modify the model, instruction, robot state, controller, or environment dynamics. The perturbation is deployed as either a translucent camera film or a physical patch in the workspace.

3 Method

3.1 Overview

We propose Causal Attack via Gradient Geometry (CAGE), a gradient-based framework that attacks WAMs by manipulating imagination–action causal consistency. We optimize differentiable imagination and action surrogates with tailored gradient strategies for disruptive and hijacking manipulation. The overall framework is illustrated in Fig. 1.

3.2 Manipulating Imagination–Action Causal Consistency

Causal objective and differentiable surrogates. As defined in Sec. 2, D_{cau} measures the discrepancy between the imagined future and the future induced by the generated actions. Since its evaluation requires environment rollouts outside the differentiable computation graph, D_{cau} cannot directly provide gradients with respect to the perturbation. We therefore introduce differentiable branch-wise objectives:

$$D_V(V_1, V_2) = \frac{1}{N_V}\left\|M_V \odot (\overline{V}_1 - \overline{V}_2)\right\|_F^2, \qquad D_A(A_1, A_2) = \frac{1}{N_A}\left\|M_A \odot (\overline{A}_1 - \overline{A}_2)\right\|_F^2. \tag{8}$$

where $\overline{V}$ and $\overline{A}$ denote normalized future representations and action sequences, M_V and M_A mask invalid elements, and N_V and N_A denote the corresponding numbers of valid entries. These objectives provide differentiable signals for indirectly manipulating the causal relationship between imagination and action.

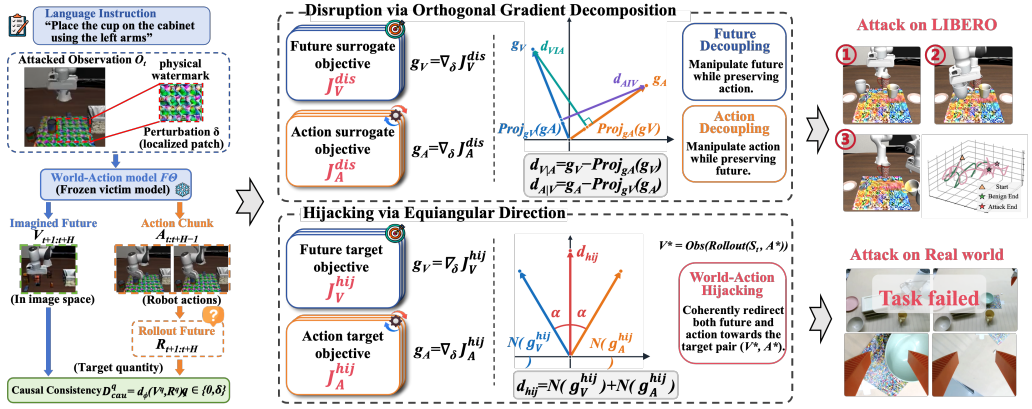


Figure 1: Overview of the proposed CAGE framework.

Disruption via orthogonal gradient decomposition. We employ orthogonal gradient decomposition to selectively manipulate the responses of the imagination and action branches. Specifically, one branch is driven away from its clean output while the other is kept locally invariant, inducing a disruption of their causal consistency. We define the branch-wise surrogate objectives and their patch gradients as

$$\mathcal{J}_V^{\text{dis}} = \mathbb{E}[D_V(V^\delta, V^0)], \quad \mathcal{J}_A^{\text{dis}} = \mathbb{E}[D_A(A^\delta, A^0)], \quad g_V^{\text{dis}} = \nabla_\delta \mathcal{J}_V^{\text{dis}}, \quad g_A^{\text{dis}} = \nabla_\delta \mathcal{J}_A^{\text{dis}}. \tag{9}$$

Since both branches respond to the same perturbation, directly optimizing either objective may simultaneously affect the other. We therefore isolate the optimization signal associated with one branch by removing its component along the gradient of the branch to be preserved Désidéri (2012); Sener & Koltun (2018); Yu et al. (2020); Liu et al. (2021). Let

$$\text{Proj}_b(a) = \frac{\langle a, b\rangle}{\|b\|_2^2 + \epsilon}\, b \tag{10}$$

denote the projection of a onto b. The two disruption directions are then

$$d_{V|A} = g_V^{\text{dis}} - \text{Proj}_{g_A^{\text{dis}}}\left(g_V^{\text{dis}}\right), \qquad d_{A|V} = g_A^{\text{dis}} - \text{Proj}_{g_V^{\text{dis}}}\left(g_A^{\text{dis}}\right). \tag{11}$$

Under exact projection, each direction maximizes the first-order increase of the targeted deviation objective subject to zero first-order change in the other deviation objective. This selectivity concerns the scalar surrogate objectives and does not imply that the preserved branch output is pointwise invariant. The numerical stabilization used in implementation makes the orthogonality approximate, as analyzed in Appendix B. In this way, orthogonal gradient decomposition enables selective manipulation of the two branches and induces disruptive changes in imagination–action causal consistency.

Hijacking via equiangular direction. We employ an equiangular direction to jointly manipulate the imagination and action branches toward a causally matched adversarial behavior. We first construct the target pair by executing a target action chunk from the same environment state:

$$O^\star_{t+1:t+H} = \text{Obs}\big(\text{Rollout}\left(S_t, A^\star_{t:t+H-1}\right)\big), \qquad V^\star = E_\phi(O^\star). \tag{12}$$

The pair $(V^\star, A^\star)$ therefore represents the same adversarial behavior from the imagination and action perspectives. We define the targeted surrogate objectives and their patch gradients as

$$\mathcal{J}_V^{\text{hij}} = -\mathbb{E}[D_V(V^\delta, V^\star)], \ \mathcal{J}_A^{\text{hij}} = -\mathbb{E}[D_A(A^\delta, A^\star)], \ g_V^{\text{hij}} = \nabla_\delta \mathcal{J}_V^{\text{hij}}, \ g_A^{\text{hij}} = \nabla_\delta \mathcal{J}_A^{\text{hij}}. \tag{13}$$

Since both branches need to be coordinated toward the causally matched target pair, directly combining their gradients can bias the optimization toward the objective with the larger gradient magnitude. We therefore normalize the gradients and construct a balanced update direction:

$$\mathcal{N}(g) = \frac{g}{\|g\|_2 + \epsilon}, \qquad d_{\text{hij}} = \mathcal{N}\left(g_V^{\text{hij}}\right) + \mathcal{N}\left(g_A^{\text{hij}}\right). \tag{14}$$

For nonzero, non-opposing gradients, this direction is the angular bisector when $\epsilon = 0$; the stabilized direction in Eq. (14) approaches it as $\epsilon \to 0$. The equiangular direction itself does not enforce causal consistency; rather, it coordinates both branches toward the target pair whose causal consistency is established through environment rollout. In this way, the two branches are coherently redirected toward the same adversarial behavior, inducing a hijacking of imagination–action causal consistency.

Optimization guarantees. With exact expected gradients, $\epsilon = 0$, and inactive box projection, the proposed directions admit the following local multi-objective interpretation. For disruption, let g_T and g_P denote the gradients of the target and preserved branch objectives, respectively. When $g_P \neq 0$, the normalized orthogonal component

$$u_{T|P} = \frac{g_T - \frac{\langle g_T, g_P \rangle}{\|g_P\|_2^2} g_P}{\left\| g_T - \frac{\langle g_T, g_P \rangle}{\|g_P\|_2^2} g_P \right\|_2} \tag{15}$$

maximizes the first-order increase of the target objective among unit-norm directions that leave the preserved objective unchanged to first order. For hijacking, let u_V and u_A be the unit-normalized gradients of the imagination and action objectives, and let θ denote the angle between them. The equiangular direction achieves the maximum worst-case normalized first-order progress:

$$\frac{u_V + u_A}{\|u_V + u_A\|_2} \in \arg\max_{\|u\|_2 \leq 1} \min\{\langle u_V, u \rangle, \langle u_A, u \rangle\}, \quad \max_{\|u\|_2 \leq 1} \min\{\langle u_V, u \rangle, \langle u_A, u \rangle\} = \cos\frac{\theta}{2}, \tag{16}$$

for $\theta < \pi$. Full proofs and the corresponding regularity conditions are provided in Appendix B.

3.3 Universal Watermark Optimization and Physical Deployment

To obtain a robust and physically deployable perturbation, we optimize a semi-transparent adversarial watermark over the input observation. During optimization, the watermark is rendered as

$$\Pi_\eta(X; \delta) = (1 - \alpha M_\eta) \odot X + \alpha M_\eta \odot \mathcal{T}_\eta(\delta), \tag{17}$$

where δ is the learnable watermark, α controls its transparency, M_η specifies the valid region, and $\mathcal{T}_\eta$ applies geometric and photometric transformations. We randomize position, scale, perspective, illumination, and image quality during optimization to improve robustness.

We optimize the watermark using $d_{V|A}$, $d_{A|V}$, and d_{hij} for future decoupling, action decoupling, and hijacking manipulation, respectively. The gradients are aggregated over each minibatch and sampled transformations, followed by projected gradient ascent:

$$\delta \leftarrow \mathcal{P}_{[0,1]} \left(\delta + \gamma \frac{d}{\|d\|_2 + \epsilon} \right), \tag{18}$$

where d denotes the corresponding update direction, γ is the step size, and $\mathcal{P}_{[0,1]}$ projects the watermark onto the valid RGB range. The victim WAM remains frozen throughout optimization. For physical deployment, the optimized perturbation is printed as an adversarial patch and placed directly in the workspace.

4 Experiments

We evaluate CAGE from four perspectives: attack effectiveness on LIBERO and RoboTwin (Sec. 4.1), transferability to real-world robots (Sec. 4.2), robustness, generalization, and visualization analyses (Sec. 4.3), and resistance to representative defenses (Sec. 4.5).

4.1 Main Results

Attack Effectiveness. As shown in Table 1, explicitly manipulating imagination–action causal consistency substantially degrades task performance across all evaluated benchmarks. Compared with the average clean SR of 96.0%, Future Decoupling, Action Decoupling, and World–Action Hijacking reduce the average SR to 15.2%, 37.0%, and 28.6%, respectively. In particular, Future Decoupling and World–Action Hijacking both completely suppress task success on LIBERO-Long (0.0% SR), demonstrating the effectiveness of consistency manipulation across diverse tasks.

Beyond Generic Input Corruption. The performance degradation caused by our attacks substantially exceeds that of generic visual perturbations. Random and Gaussian perturbations retain average SRs of 89.0% and 90.4%, respectively, whereas our optimized attacks reduce the average SR to as low as 15.2%. This large gap suggests that the observed failures arise from targeted manipulation of the WAM decision process rather than generic input corruption. We further examine this distinction through feature-space analysis in Sec. 4.4.

Table 1: Main results of universal patch attacks against LingBot-VA. D_V and D_A denote imagination and action deviations, respectively; $\hat{D}_{\text{cau}}$ is a computable proxy for imagination–action causal consistency (Appendix A.2). Clean SR and SR denote task success rates without and under attack. All values are reported in percentages (%).

Attack / Setting	Metric	LIBERO-Object	LIBERO-Spatial	LIBERO-Goal	LIBERO-Long	RoboTwin	Average
Clean	SR	99.0	96.0	97.0	99.0	89.0	**96.0**
Random Perturb.	**SR** ↓	92.2	88.0	92.0	87.0	86.0	**89.0**
Gaussian Perturb.	**SR** ↓	94.2	86.0	97.0	90.0	85.0	**90.4**
Future Dec.	**SR** ↓	10.0	26.0	31.0	0.0	9.0	**15.2**
	D_V	112.8	103.4	110.1	113.7	60.3	100.1
	D_A	95.9	35.0	65.5	62.8	31.2	58.1
	$\hat{D}_{\text{cau}}$	37.2	5.4	19.5	12.9	20.5	19.1
Action Dec.	**SR** ↓	67.0	14.0	37.0	41.0	26.0	**37.0**
	D_V	74.7	85.7	78.9	72.9	56.5	73.7
	D_A	29.9	41.6	103.1	70.3	28.4	54.7
	$\hat{D}_{\text{cau}}$	42.0	3.8	23.6	49.7	24.4	28.7
W.–A. Hijack.	**SR** ↓	28.0	32.0	29.0	0.0	54.0	**28.6**
	D_V	109.6	108.9	104.4	111.7	64.2	99.8
	D_A	131.7	58.0	64.9	79.8	23.3	71.5
	$\hat{D}_{\text{cau}}$	39.6	5.0	24.4	65.5	53.4	37.6

Distinct Manipulation Patterns. The three attack modes induce distinct effects on imagination and action. Future Decoupling produces the largest imagination deviation, reaching an average D_V of 100.1, while Action Decoupling yields substantially lower imagination deviation (73.7) while maintaining considerable action deviation (54.7). World–Action Hijacking jointly perturbs both branches, with average D_V and D_A reaching 99.8 and 71.5, respectively. These results show that imagination–action causal consistency constitutes a flexible attack surface through which different manipulation patterns can be induced in the prediction and control branches.

4.2 Real-World Attack

We investigate whether CAGE transfers to real-world robotic systems. We deploy the physical patch on a dual-arm Franka Research 3 (FR3) platform with the LingBot-VA policy, without modifying the model or controller. We evaluate three real-world manipulation tasks, where the clean policy achieves success rates of $8/10$, $10/10$, and $8/10$, respectively. After applying the physical patch, the success rates decrease to $3/10$, $5/10$, and $1/10$. As shown in Figure 2, the policy initially follows similar trajectories before the patch enters the camera view. Once the robot approaches the patch, the predicted actions begin to deviate, eventually causing task failures. These results demonstrate that CAGE can effectively transfer to real-world deployment and disrupt WAM-based closed-loop control.

4.3 Diagnostic Experiments

We investigate how deployment conditions affect the behavior of CAGE, including perturbation budget and visual observation streams. Results on LIBERO-Long and RoboTwin are summarized in Tables 2 and 3.

Table 2: Sensitivity analysis on LIBERO-Long. All values are reported in percentages (%).

Factor	Setting	Clean SR	Future Decoupling		Action Decoupling		World–Action Hijacking	
			$\hat{D}_{\text{cau}}$	SR↓	$\hat{D}_{\text{cau}}$	SR↓	$\hat{D}_{\text{cau}}$	SR↓
Patch Size	50%	99.0	0.7	79.0	0.5	84.0	1.3	74.0
	70%	99.0	5.6	75.0	6.5	85.0	7.9	84.0
	90%	99.0	24.3	35.0	14.1	61.0	30.5	48.0
Opacity	16/255	99.0	39.7	12.0	14.7	81.0	152.1	90.0
	32/255	99.0	12.9	0.0	49.7	41.0	65.5	0.0
	48/255	99.0	29.6	0.0	119.5	15.0	64.9	0.0
Camera	Main view	99.0	14.0	58.0	5.4	55.0	15.7	93.0
	Wrist view	99.0	12.9	0.0	49.7	41.0	65.5	0.0
	Dual view	99.0	5.5	0.0	56.5	41.0	27.1	0.0

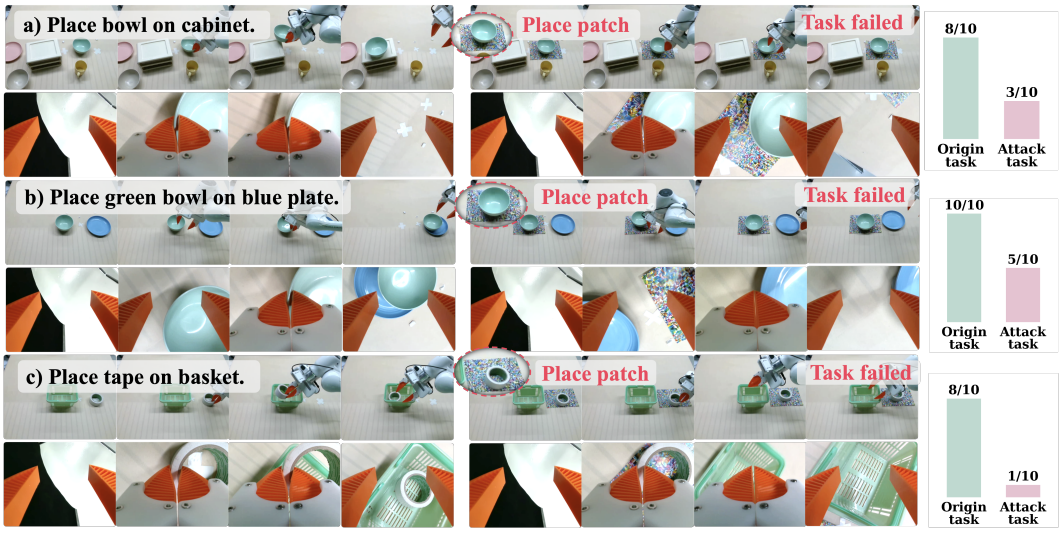


Figure 2: Real-world transferability of CAGE on a Franka Research 3 (FR3) robotic platform.

Table 3: Sensitivity analysis on RoboTwin. All values are reported in percentages (%).

Factor	Setting	Clean SR	Future Decoupling		Action Decoupling		World–Action Hijacking	
			$\hat{D}_{\text{cau}}$	SR↓	$\hat{D}_{\text{cau}}$	SR↓	$\hat{D}_{\text{cau}}$	SR↓
Patch Size	50%	89.0	0.3	78.0	1.4	73.0	1.7	62.0
	70%	89.0	5.0	64.0	5.8	68.0	9.8	72.0
	90%	89.0	7.3	30.0	3.7	24.0	5.0	56.0
Opacity	16/255	89.0	4.3	41.0	3.6	66.0	11.5	61.0
	32/255	89.0	20.5	9.0	24.4	26.0	5.6	54.0
	48/255	89.0	16.6	4.0	9.6	20.0	9.3	15.0
Camera	Main view	89.0	5.6	67.0	5.6	40.0	2.6	53.0
	Wrist view	89.0	20.5	9.0	24.4	26.0	5.6	54.0
	Dual view	89.0	6.6	9.0	9.5	26.0	15.8	54.0

Sensitivity to perturbation budget. We first analyze the impact of patch size and opacity. Attack effectiveness generally improves with increasing perturbation budget. For example, on RoboTwin, increasing opacity from $16/255$ to $32/255$ reduces the SR of Future Decoupling from 41.0% to 9.0%, while a further increase to $48/255$ reduces it to 4.0%. A similar trend is observed for patch size, although the effect is not strictly monotonic across all attack modes.

Sensitivity to visual observation streams. We further examine the effect of camera configurations. Wrist-view attacks are generally more effective than main-view attacks. On LIBERO-Long, wrist-view perturbations reduce SR to 0.0% for both Future Decoupling and World–Action Hijacking, compared with 58.0% and 93.0% under the main view. This difference likely arises because the wrist camera provides a closer view of the manipulation region, making local perturbations more prominent in task-relevant visual observations.

4.4 Visualization Analysis

Impact on Imagined Futures. Figure 3 qualitatively compares imagined futures under clean and adversarial inputs. While clean predictions maintain recognizable objects and coherent task progression, attacked predictions exhibit substantial changes in object appearance and spatial configuration, including object displacement and distortion that become more pronounced in later frames. These visualizations help explain the substantial drop in task success rate under adversarial perturbations reported in Table 1.

Trajectory Analysis. To better understand the effect of CAGE, we visualize clean and attacked trajectories in Fig. 4. The trajectories initially remain close, when the adversarial patch has limited



Figure 3: Imagined futures under clean and adversarial inputs.

influence on the robot's observation. As execution proceeds, the attacked trajectory progressively diverges from the clean trajectory and exhibits irregular motions, eventually driving the robot away from the intended execution path and causing unintended interactions with the environment.

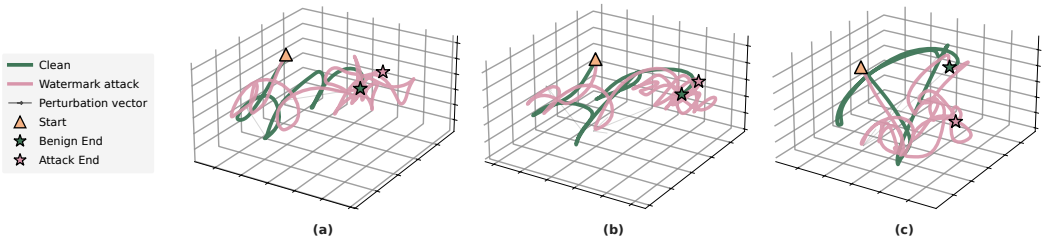


Figure 4: Attack trajectories of CAGE on LIBERO. The panels show (a) a targeted attack and (b) an untargeted attack on the mug-placement task, and (c) an untargeted attack on the basket-placement task. Dark green and light pink denote clean and attacked trajectories, respectively.

Feature Analysis. To further analyze how CAGE affects the model's perception, we visualize the PCA distributions of perceptual features under clean, Gaussian-noise, and watermark-attack conditions in Fig. 5. Compared with Gaussian noise, the optimized watermark consistently shifts the perceptual features farther away from their clean distributions across different tasks. This larger feature-space deviation indicates that CAGE more effectively disrupts the model's internal perception, providing further evidence for the effectiveness of the optimized adversarial watermark.

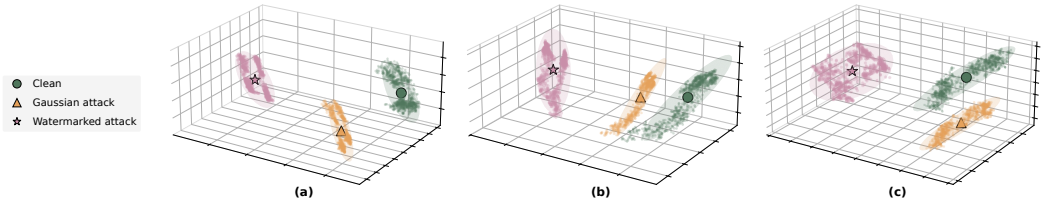


Figure 5: Three-dimensional PCA feature distributions. Each panel compares clean, Gaussian-noise, and watermark-attack conditions, shown in dark green, orange, and light pink, respectively.

4.5 Defense Experiments

We evaluate whether existing defense methods can mitigate CAGE without retraining the victim WAM. Specifically, we consider JPEG compression with quality $q = 80$, bit-depth reduction to $b = 4$ bits, median filtering with a 3×3 kernel Xu et al. (2018), and random resize with padding at a 90% resize ratio Xie et al. (2018). As shown in Table 4, these defenses provide limited mitigation against our attacks. For example, on LIBERO-Long, Future Decoupling remains highly effective after all defenses, with SRs of 5.0%, 1.0%, 1.0%, and 0.0%, compared with 0.0% without defense. Similarly, World–Action Hijacking remains largely unaffected, with SRs close to zero across different defense methods. Although some defenses partially improve SR for Action Decoupling, no defense method consistently mitigates all attack modes. These results uncover a new attack surface in WAMs caused by the vulnerability of imagination–action causal consistency. As CAGE disrupts the alignment

between imagined futures and executed actions, conventional input-level defenses are insufficient, highlighting the need for dedicated defense strategies against this emerging threat.

Table 4: Defense results on LIBERO-Long and RoboTwin.

(a) LIBERO-Long

Defense	Clean SR	Future Decoupling		Action Decoupling		World–Action Hijacking	
		$\hat{D_{\mathrm{cau}}}$	SR	$\hat{D_{\mathrm{cau}}}$	SR	$\hat{D_{\mathrm{cau}}}$	SR
No Defense	99.0	12.9	0.0	49.7	41.0	43.8	0.0
JPEG Compression	98.0	11.1	5.0	28.6	63.0	43.5	1.0
Bit-depth Reduction	98.0	12.7	1.0	37.6	52.0	44.3	1.0
Median Filtering	98.0	12.1	1.0	33.2	56.0	44.4	2.0
Random Resize + Padding	76.0	13.0	0.0	53.7	0.0	41.2	0.0

(b) RoboTwin

Defense	Clean SR	Future Decoupling		Action Decoupling		World–Action Hijacking	
		$\hat{D_{\mathrm{cau}}}$	SR	$\hat{D_{\mathrm{cau}}}$	SR	$\hat{D_{\mathrm{cau}}}$	SR
No Defense	89.0	20.5	9.0	24.4	26.0	5.6	54.0
JPEG Compression	88.0	10.9	9.0	13.0	32.0	6.2	56.0
Bit-depth Reduction	89.0	9.7	15.0	13.5	24.0	5.5	55.0
Median Filtering	88.0	10.1	16.0	12.9	24.0	5.7	58.0
Random Resize + Padding	84.0	11.6	17.0	12.4	25.0	5.8	54.0

5 Related Work

World-Action Models and Embodied Attacks. World-model-based control leverages predictive representations for decision making and robot planning Ha & Schmidhuber (2018); Hafner et al. (2020); Finn & Levine (2017). Building on generalist VLA policies Zitkovich et al. (2023); Kim et al. (2025); Ghosh et al. (2024), recent WAMs jointly model future observations and actions Du et al. (2023); Tian et al. (2024); Hu et al. (2024); Li et al. (2025b); Gao et al. (2025); Assran et al. (2025); Ye et al. (2026); Wang et al. (2026b); Li et al. (2026a), with growing attention to imagination–action consistency Ruan et al. (2026); Wang et al. (2026a) and security Liu et al. (2026); Chen et al. (2026); Li et al. (2026b); Shen et al. (2026). Meanwhile, adversarial attacks have progressed from digital and physical perturbations Szegedy et al. (2014); Goodfellow et al. (2014); Brown et al. (2017); Athalye et al. (2018); Moosavi-Dezfooli et al. (2017) and RL agents Huang et al. (2017); Lin et al. (2017); Kos & Song (2017); Zhang et al. (2020) to embodied systems Melis et al. (2017); Wang et al. (2019); Cheng et al. (2022); Abdelfattah et al. (2021); Islam et al. (2024); Chen et al. (2025), including attacks on embodied LLMs and VLAs Liu et al. (2024); Zhang et al. (2024); Jiao et al. (2025); Wang et al. (2024; 2025a); Jones et al. (2025); Wang et al. (2025b); Zhou et al. (2025); Li et al. (2025a); Wang et al. (2026c), transferable physical patches Lu et al. (2026), and reasoning attacks Huang et al. (2026). We target the causal coupling between WAM imagination and action, enabling both disruption and hijacking of their consistency.

6 Conclusion

In this paper, we reveal a security vulnerability in the imagination–action causal consistency of World-Action Models and propose a gradient-based attack framework Causal Attack via Gradient Geometry that enables two attack modes: disruption and hijacking. Extensive experiments in both simulated and real-world robotic environments demonstrate the effectiveness and transferability of our attacks. Furthermore, defense evaluations show that existing input-level preprocessing techniques provide limited mitigation against such attacks, highlighting the need for robust training, verification, and defense mechanisms. We hope this work motivates further research toward building secure and reliable multimodal robot policies for next-generation embodied intelligence systems.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing and code development. Specifically, they were used to improve the clarity, grammar, and presentation of author-written text, and to assist with implementing, debugging, and refining experimental and visualization code. All AI-assisted content was subject to human supervision and review by the authors. The authors manually examined and revised the generated text, and verified and tested AI-assisted code before use. Generative AI was not used as a substitute for scientific judgment, and the research questions, methodology, experimental design, analysis, and conclusions were determined and validated by the authors. We take full responsibility for the final content of this work, including any text or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work investigates adversarial vulnerabilities in world–action models with the goal of improving the security and reliability of embodied AI systems. The proposed attacks are developed and evaluated solely for research purposes in controlled simulated and laboratory environments. Real-world experiments are conducted on research robotic platforms under human supervision, without involving human subjects or personal data. While the proposed techniques could potentially be misused to interfere with deployed robotic systems, our study is intended to expose previously unexplored risks in imagination–action causal consistency and motivate the development of corresponding defenses. All experiments follow controlled evaluation protocols, and we encourage responsible use of the methods and findings presented in this work.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed attack framework, optimization objectives, and implementation procedures in the main paper and appendix. The appendix further documents the experimental settings, hyperparameters, perturbation configurations, evaluation protocols, and additional results across LIBERO, RoboTwin, and real-world robotic experiments. All experiments follow the configurations and evaluation procedures described in the paper. We will release the implementation code and necessary evaluation scripts to facilitate reproduction of the reported results.

REFERENCES

Mazen Abdelfattah et al. Adversarial attacks on camera-LiDAR models for 3d car detection. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2189–2194, 2021. doi: 10.1109/IROS51168.2021.9636638.

Mahmoud Assran et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning, 2025.

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 284–293, 2018.

Tom B. Brown, Dandelion Mané, Aurko Roy, Martin Abadi, and Justin Gilmer. Adversarial patch, 2017.

Linghan Chen, Kaiyan Ji, and Minyu Guo. Attacking the trusted imagination: Oracle-level integrity attacks on imagine-then-act world models, 2026.

Meng Chen et al. Towards physically realizable adversarial attacks in embodied vision navigation. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 11819–11825, 2025. doi: 10.1109/IROS60139.2025.11246438.

Zhiyuan Cheng et al. Physical attack on monocular depth estimation with optimal adversarial patches, 2022.

Jean-Antoine Désidéri. Multiple-gradient descent algorithm (MGDA) for multiobjective optimization. *Comptes Rendus Mathématique*, 350(5–6):313–318, 2012. doi: 10.1016/j.crma.2012.03.014.

Yilun Du et al. Learning universal policies via text-guided video generation, 2023.

Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *2017 IEEE International Conference on Robotics and Automation*, pp. 2786–2793, 2017. doi: 10.1109/ICRA.2017.7989324.

Shenyuan Gao et al. Adaworld: Learning adaptable world models with latent actions, 2025.

Dibya Ghosh, Homer Rich Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Quan Vuong, Ted Xiao, Pannag R. Sanketi, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, 2024. doi: 10.15607/RSS.2024.XX.090.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2014.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

Yucheng Hu et al. Video prediction policy: A generalist robot policy with predictive visual representations, 2024.

Sandy H. Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial attacks on neural network policies, 2017.

Zhengxian Huang, Wenjun Zhu, Haoxuan Qiu, Xiaoyu Ji, and Wenyuan Xu. TRAP: Hijacking VLA CoT-reasoning via adversarial patches. *arXiv preprint arXiv:2603.23117*, 2026.

Chashi Mahiul Islam et al. Malicious path manipulations via exploitation of representation vulnerabilities of vision-language navigation systems. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 13845–13852, 2024. doi: 10.1109/IROS58592.2024.10802618.

Ruochen Jiao, Shaoyuan Xie, Justin Yue, Takami Sato, Lixu Wang, Yixuan Wang, Qi Alfred Chen, and Qi Zhu. Can we trust embodied agents? exploring backdoor attacks against embodied LLM-based decision-making systems. In *International Conference on Learning Representations*, 2025.

Eliot Jones, Alexander Robey, Andy Zou, et al. Adversarial attacks on robotic vision-language-action models, 2025.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P. Foster, Pannag R. Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. In *Proceedings of the 8th Conference on Robot Learning*, pp. 2679–2713, 2025.

Jernej Kos and Dawn Song. Delving into adversarial attacks on deep policies, 2017.

Jiayu Li et al. AttackVLA: Benchmarking adversarial and backdoor attacks on vision-language-action models, 2025a.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, et al. Causal world modeling for robot control, 2026a.

Qi Li, Xingyi Yang, and Xinchao Wang. BadWAM: When world-action models dream right but act wrong, 2026b.

Shuang Li et al. Unified video action model, 2025b.

Yen-Chen Lin et al. Tactics of adversarial attack on deep reinforcement learning agents. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 3756–3762, 2017. doi: 10.24963/ijcai.2017/525.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18878–18890, 2021.

Hanqing Liu et al. JailWAM: Jailbreaking world action models in robot control, 2026.

Shuyuan Liu et al. Exploring the robustness of decision-level through adversarial attacks on LLM-based embodied models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024. doi: 10.1145/3664647.3680616.

Hui Lu, Yi Yu, Yiming Yang, Chenyu Yi, Qixin Zhang, Bingquan Shen, Alex C. Kot, and Xudong Jiang. When robots obey the patch: Universal transferable patch attacks on vision-language-action models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22867–22878, 2026.

Marco Melis et al. Is deep learning safe for robot vision? adversarial examples against the iCub humanoid. In *2017 IEEE International Conference on Computer Vision Workshops*, pp. 751–759, 2017. doi: 10.1109/ICCVW.2017.94.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1765–1773, 2017.

Bo-Kai Ruan, Teng-Fang Hsiao, Ling Lo, and Hong-Han Shuai. Is the future compatible? diagnosing dynamic consistency in world action models, 2026.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, volume 31, pp. 527–538, 2018.

Linghui Shen, Mingyue Cui, and Xingyi Yang. BadWorld: Adversarial attacks on world models, 2026.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Yang Tian et al. Predictive inverse dynamics models are scalable learners for robotic manipulation, 2024.

David Wang et al. Adversarial grasp objects. In *2019 IEEE 15th International Conference on Automation Science and Engineering*, pp. 241–248, 2019. doi: 10.1109/COASE.2019.8843059.

Rui Wang, Yue Zhang, Jiehong Lin, Kuncheng Luo, Jianan Wang, Zhongrui Wang, and Xiaojuan Qi. When to trust imagination: Adaptive action execution for world action models, 2026a.

Siyin Wang et al. World action models: The next frontier in embodied ai, 2026b.

Taowen Wang, Cheng Han, James Liang, Wenhao Yang, Dongfang Liu, Luna Xinyu Zhang, Qifan Wang, Jiebo Luo, and Ruixiang Tang. Exploring the adversarial vulnerabilities of vision-language-action models in robotics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6948–6958, 2025a. doi: 10.1109/ICCV51701.2025.00653.

Xianlong Wang et al. TrojanRobot: Backdoor attacks against robotic manipulation in the physical world, 2024.

Xiaofei Wang, Mingliang Han, Tianyu Hao, Yi Yang, Yun-Bo Zhao, and Keke Tang. Partially observable adversarial patch attacks on vision-language-action models in robotics. *IEEE Robotics and Automation Letters*, pp. 9351–9358, 2026c. doi: 10.1109/LRA.2026.3703262.

Yichen Wang et al. ADVEDM: Fine-grained adversarial attack against VLM-based embodied agents, 2025b.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan L. Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations*, 2018.

Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *Network and Distributed System Security Symposium*, 2018. doi: 10.14722/ndss.2018.23198.

Seonghyeon Ye et al. World action models are zero-shot policies, 2026.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5824–5836, 2020.

Hangtao Zhang et al. BadRobot: Jailbreaking embodied LLM agents in the physical world, 2024. Published at ICLR 2025.

Huan Zhang et al. Robust deep reinforcement learning against adversarial perturbations on state observations, 2020.

Xueyang Zhou et al. BadVLA: Towards backdoor attacks on vision-language-action models via objective-decoupled optimization, 2025.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of the 7th Conference on Robot Learning*, pp. 2165–2183, 2023.

A Appendix

A.1 Real-World Experimental Setup

The real-world experiments are conducted on a dual-arm robotic platform consisting of two Franka Research 3 seven-degree-of-freedom robotic arms. Each arm is equipped with a Franka Hand and custom 3D-printed fingers. The vision system consists of three Intel RealSense D435i cameras. Two cameras are mounted on the end-effectors of the left and right robotic arms, respectively, to provide local manipulation views, while the third camera is used to capture a global view of the workspace.

The watermark used in the real-world experiments is obtained through offline optimization, printed on an A4 sheet of paper, and placed underneath the target objects on the tabletop. During watermark optimization, all parameters of LingBot-VA are frozen, and only the pixel values of the universal watermark are updated. Three robotic manipulation tasks are selected for the experiments, and their language instructions and training/validation splits are shown in Table 5. Each task contains 40 training trajectories and 10 validation trajectories, resulting in a total of 120 training trajectories and 30 validation trajectories. For each training trajectory, 64 frames are uniformly sampled as optimization samples, and the batch size is set to 4. The watermark is optimized for 10 epochs, corresponding to 19,200 iterative updates. The final watermark has a resolution of 224×320, and the blending coefficient is set to $\alpha = 32/255$.

A.2 Computable Proxy for Imagination–Action Causal Consistency

Imagination–action causal consistency cannot be directly quantified because it requires measuring whether the imagined future is causally compatible with the future induced by the predicted actions. We therefore construct a computable proxy, $\hat{D}_{\mathrm{cau}}$, by comparing motion in the model-predicted

Table 5: Task Instructions and Data Splits for Watermark Training

Task	Instruction	Train	Val.
1	Place the green bowl on top of the cabinet with the left arm.	40	10
2	Place the green bowl on the blue plate with the left arm.	40	10
3	Place the tape in the green basket with the left arm.	40	10
Total		120	30

video with motion observed after executing the corresponding predicted actions. For each episode, clean and adversarial rollouts start from the same initial environment state. At each decision step, the model predicts a future video segment and an action sequence. We execute the action sequence in the simulator and record the resulting observation video. Gripper surface points projected from the simulator are used as shared query points, which are independently tracked in the predicted and executed videos using CoTracker3. For each tracked point, motion is represented by its displacement relative to the beginning of the decision step. We compute the discrepancy between the normalized displacement trajectories in the predicted and executed videos and aggregate it over tracked points and temporal positions. Visibility and tracking confidence from both videos are used as weights to reduce the influence of occlusion and unreliable tracks.

For LIBERO, we use gripper trajectories from the wrist camera. For RoboTwin, discrepancies from the left and right wrist views are computed separately and equally averaged. Let D_{clean} and D_{attack} denote the resulting motion discrepancies under clean and adversarial conditions. We define the attack-induced relative change as

$$\hat{D}_{\text{cau}} = \frac{D_{\text{attack}} - D_{\text{clean}}}{D_{\text{clean}}} \times 100\%. \tag{19}$$

A larger $\hat{D}_{\text{cau}}$ indicates a greater attack-induced mismatch between the imagined future and the future induced by the predicted actions, and therefore stronger disruption of imagination–action causal consistency.

B Theoretical Analysis

B.1 Notation and Scope

All gradient results in this section concern the differentiable branch-wise objectives with respect to the perturbation δ. Unless stated otherwise, gradients are exact gradients of the corresponding expected objectives. The finite-step results apply to updates for which the line segment between the current and updated perturbations lies in the stated local neighborhood and the box projection is inactive. Minibatch gradient estimation and clipping may change the realized update direction.

For a differentiable objective J, we write $g = \nabla_\delta J(\delta)$. A function J is locally L-smooth if its gradient is L-Lipschitz in the considered neighborhood. In this case, for every admissible displacement h,

$$|J(\delta + h) - J(\delta) - \langle \nabla J(\delta), h \rangle| \le \frac{L}{2}\|h\|_2^2. \tag{20}$$

We first analyze the exact projection and exact gradient normalization. The effects of the numerical constant ϵ used in the implementation are addressed in Appendix B.5.

B.2 Orthogonal Gradient Decomposition

Let J_T and J_P denote the objectives of the target branch and the branch to be preserved, respectively, and set $g_T = \nabla_\delta J_T$ and $g_P = \nabla_\delta J_P$. Suppose $g_P \neq 0$, and define

$$d_{T|P} = g_T - \frac{\langle g_T, g_P \rangle}{\|g_P\|_2^2} g_P. \tag{21}$$

Proposition B.1 (Selective first-order optimality). If $d_{T|P} \neq 0$, then $u_{T|P} = d_{T|P}/\|d_{T|P}\|_2$ solves

$$\max_{\|u\|_2 \leq 1} \langle g_T, u \rangle \quad \text{subject to} \quad \langle g_P, u \rangle = 0. \tag{22}$$

If θ is the angle between g_T and g_P, the optimal value is $\|g_T\|_2 |\sin\theta|$.

Proof. By construction,

$$\langle g_P, d_{T|P} \rangle = \langle g_P, g_T \rangle - \frac{\langle g_T, g_P \rangle}{\|g_P\|_2^2} \|g_P\|_2^2 = 0. \tag{23}$$

Thus $u_{T|P}$ satisfies the constraint in Eq. equation 22. For any feasible u, the decomposition

$$g_T = d_{T|P} + \frac{\langle g_T, g_P \rangle}{\|g_P\|_2^2} g_P$$

and the constraint $\langle g_P, u \rangle = 0$ imply

$$\langle g_T, u \rangle = \langle d_{T|P}, u \rangle \leq \|d_{T|P}\|_2 \|u\|_2 \leq \|d_{T|P}\|_2. \tag{24}$$

Both inequalities are equalities for $u = u_{T|P}$. Finally,

$$\|d_{T|P}\|_2^2 = \|g_T\|_2^2 - \frac{\langle g_T, g_P \rangle^2}{\|g_P\|_2^2} \tag{25}$$

$$= \|g_T\|_2^2 (1 - \cos^2\theta) = \|g_T\|_2^2 \sin^2\theta. \tag{26}$$

Taking square roots proves the stated optimal value. □

Corollary B.1 (Finite-step selectivity). Suppose J_T and J_P are locally L_T-smooth and L_P-smooth, respectively. For an admissible update $\delta' = \delta + \gamma u_{T|P}$,

$$J_T(\delta') - J_T(\delta) \geq \gamma \|d_{T|P}\|_2 - \frac{L_T}{2}\gamma^2, \tag{27}$$

$$|J_P(\delta') - J_P(\delta)| \leq \frac{L_P}{2}\gamma^2. \tag{28}$$

Proof. Proposition B.1 gives $\langle g_T, u_{T|P} \rangle = \|d_{T|P}\|_2$ and $\langle g_P, u_{T|P} \rangle = 0$. Apply Eq. equation 20 to each objective with $h = \gamma u_{T|P}$ and $\|u_{T|P}\|_2 = 1$. □

In particular, the lower bound in Eq. equation 27 is positive when $0 < \gamma < 2\|d_{T|P}\|_2/L_T$, provided $L_T > 0$. If $g_P = 0$, the preservation constraint is inactive and a normalized g_T is optimal when $g_T \neq 0$. If $d_{T|P} = 0$, the optimal value in Eq. equation 22 is zero: there is no feasible direction with positive first-order target progress.

B.3 Equiangular Direction for Hijacking

Let $g_V = \nabla_\delta J_V^{\text{hij}}$ and $g_A = \nabla_\delta J_A^{\text{hij}}$ be nonzero, and define

$$u_V = \frac{g_V}{\|g_V\|_2}, \qquad u_A = \frac{g_A}{\|g_A\|_2}. \tag{29}$$

Let $\theta \in [0, \pi)$ be the angle between u_V and u_A, and let

$$u_{\text{hij}} = \frac{u_V + u_A}{\|u_V + u_A\|_2}. \tag{30}$$

Proposition B.2 (Max–min first-order progress). The equiangular direction solves

$$\max_{\|u\|_2 \leq 1} \min\{\langle u_V, u \rangle, \langle u_A, u \rangle\}, \tag{31}$$

and the optimal value is $\cos(\theta/2)$.

Proof. For every feasible u,

$$\min\{\langle u_V, u\rangle, \langle u_A, u\rangle\} \le \frac{\langle u_V + u_A, u\rangle}{2} \tag{32}$$

$$\le \frac{\|u_V + u_A\|_2 \|u\|_2}{2} \tag{33}$$

$$\le \frac{\|u_V + u_A\|_2}{2}. \tag{34}$$

Because u_V and u_A are unit vectors,

$$\|u_V + u_A\|_2^2 = 2 + 2\cos\theta = 4\cos^2(\theta/2). \tag{35}$$

The upper bound is therefore $\cos(\theta/2)$. For the proposed direction,

$$\langle u_V, u_{\text{hij}}\rangle = \frac{1 + \cos\theta}{2\cos(\theta/2)} = \cos(\theta/2). \tag{36}$$

The same equality holds for u_A. Thus u_{hij} attains the upper bound and solves Eq. equation 31. □

Corollary B.2 (Simultaneous local progress). If J_i^{hij} is locally L_i-smooth for $i \in \{V, A\}$, then an admissible update $\delta' = \delta + \gamma u_{\text{hij}}$ satisfies

$$J_i^{\text{hij}}(\delta') - J_i^{\text{hij}}(\delta) \ge \gamma \|g_i\|_2 \cos(\theta/2) - \frac{L_i}{2}\gamma^2. \tag{37}$$

This follows directly from Eq. equation 20 and Proposition B.2. When $L_i > 0$, both lower bounds are positive if

$$0 < \gamma < \min_{i \in \{V,A\}} \frac{2\|g_i\|_2 \cos(\theta/2)}{L_i}. \tag{38}$$

At $\theta = \pi$, the two unit gradients are opposite. The optimal value of Eq. equation 31 is zero, and no direction yields strictly positive first-order progress for both objectives.

B.4 Relation to Rollout-Based Causal Consistency

The preceding propositions characterize the optimization of branch-wise surrogate objectives. We next state the additional conditions under which these objectives can be related to the rollout-based causal discrepancy.

For a fixed environment state S_t, let $\mathcal{R}_{S_t}(A)$ denote the observation generated by executing A from S_t, expressed in the same representation space as the imagined future. Define

$$D_{\text{cau}}(V, A) = d_\phi(V, \mathcal{R}_{S_t}(A)), \tag{39}$$

and, using the branch-wise distances defined in the main text,

$$\rho_V(V_1, V_2) = \sqrt{N_V D_V(V_1, V_2)}, \qquad \rho_A(A_1, A_2) = \sqrt{N_A D_A(A_1, A_2)}. \tag{40}$$

Assumption B.1 (Local stability). On a region containing all representations appearing in the corresponding bounds, d_ϕ is a symmetric pseudometric; in particular, it satisfies the triangle inequality and $d_\phi(x, y) = d_\phi(y, x)$. There are finite constants $L_V, L_A > 0$ such that

$$d_\phi(V_1, V_2) \le L_V \rho_V(V_1, V_2), \tag{41}$$

$$d_\phi(\mathcal{R}_{S_t}(A_1), \mathcal{R}_{S_t}(A_2)) \le L_A \rho_A(A_1, A_2). \tag{42}$$

These conditions are local assumptions on the chosen representation and the action-induced rollout; differentiability of the environment is not required.

Proposition B.3 (Hijacking surrogate bound). Let $(V^\star, A^\star)$ be the target pair and write $r^\star = D_{\mathrm{cau}}(V^\star, A^\star)$. Under Assumption B.1, every pair (V^δ, A^δ) in the specified neighborhood satisfies

$$D_{\mathrm{cau}}(V^\delta, A^\delta) \le r^\star + L_V\sqrt{N_V D_V(V^\delta, V^\star)} + L_A\sqrt{N_A D_A(A^\delta, A^\star)}. \tag{43}$$

Proof. Applying the triangle inequality through the target pair yields

$$d_\phi\big(V^\delta, \mathcal{R}_{S_t}(A^\delta)\big) \le d_\phi(V^\delta, V^\star) + d_\phi(V^\star, \mathcal{R}_{S_t}(A^\star)) + d_\phi\big(\mathcal{R}_{S_t}(A^\star), \mathcal{R}_{S_t}(A^\delta)\big). \tag{44}$$

The middle term is $r^\star$. Bound the first term using Eq. equation 41 and the last term using Eq. equation 42; then substitute the definitions of ρ_V and ρ_A. □

If the target imagination is exactly the representation of the target rollout, then $r^\star = 0$. Otherwise, the residual $r^\star$ remains in the bound. Equation equation 43 shows that reducing both surrogate distances tightens an upper bound on D_{cau}; it does not assert a monotonic decrease of the actual causal discrepancy at every update.

Conditional bounds for disruption. A change in one branch does not, on its own, imply an increase in D_{cau}: the environment rollout may change in a compensating manner. The following bounds make the sufficient conditions explicit. Let $D^0 = D_{\mathrm{cau}}(V^0, A^0)$ and $D^\delta = D_{\mathrm{cau}}(V^\delta, A^\delta)$.

For future decoupling, additionally suppose that, on the considered pairs, there is $m_V > 0$ such that

$$d_\phi(V^\delta, V^0) \ge m_V \rho_V(V^\delta, V^0). \tag{45}$$

The triangle inequality gives

$$d_\phi(V^\delta, V^0) \le D^\delta + d_\phi\big(\mathcal{R}_{S_t}(A^\delta), \mathcal{R}_{S_t}(A^0)\big) + D^0. \tag{46}$$

Combining this inequality with Eqs. equation 42 and equation 45 yields

$$D^\delta - D^0 \ge m_V \rho_V(V^\delta, V^0) - L_A \rho_A(A^\delta, A^0) - 2D^0. \tag{47}$$

Thus a sufficiently large imagination deviation relative to the action deviation and clean causal error is sufficient to increase the causal discrepancy.

For action decoupling, an analogous statement needs a lower sensitivity condition on the rollout. Suppose that there is $m_A > 0$ such that

$$d_\phi\big(\mathcal{R}_{S_t}(A^\delta), \mathcal{R}_{S_t}(A^0)\big) \ge m_A \rho_A(A^\delta, A^0) \tag{48}$$

on the considered action pairs. The triangle inequality gives

$$d_\phi\big(\mathcal{R}_{S_t}(A^\delta), \mathcal{R}_{S_t}(A^0)\big) \le D^\delta + d_\phi(V^\delta, V^0) + D^0. \tag{49}$$

Using Eqs. equation 41 and equation 48, we obtain

$$D^\delta - D^0 \ge m_A \rho_A(A^\delta, A^0) - L_V \rho_V(V^\delta, V^0) - 2D^0. \tag{50}$$

The lower sensitivity condition in Eq. equation 48 is needed because different action sequences can otherwise produce nearly identical observed futures. These conditional bounds describe when branch separation implies causal separation; they do not establish that the optimization procedure necessarily reaches those conditions.

B.5 Effect of Numerical Stabilization

The implemented orthogonal direction replaces the exact projection denominator by $\|g_P\|_2^2 + \epsilon$:

$$d^\epsilon_{T|P} = g_T - \frac{\langle g_T, g_P\rangle}{\|g_P\|_2^2 + \epsilon} g_P. \tag{51}$$

Its residual first-order interaction with the preserved objective is exactly

$$\langle g_P, d^{\epsilon}_{T|P} \rangle = \frac{\epsilon}{\|g_P\|_2^2 + \epsilon} \langle g_T, g_P \rangle. \tag{52}$$

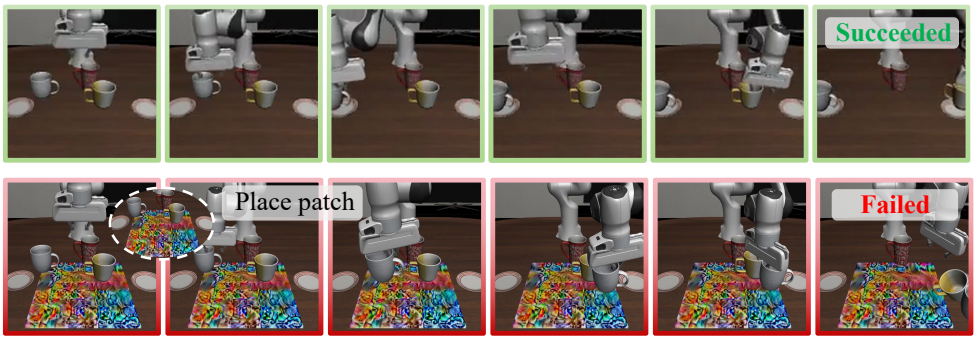
When $g_P \neq 0$, $d^{\epsilon}_{T|P}$ converges to the exact projected direction as $\epsilon \to 0$. The normalized implemented direction converges to $u_{T|P}$ if, in addition, $d_{T|P} \neq 0$. Near $d_{T|P} = 0$, normalization can amplify the residual in Eq. equation 52; hence the exact first-order preservation guarantee does not apply uniformly near this degeneracy.

For a nonzero gradient g, the stabilized normalization $N_\epsilon(g) = g/(\|g\|_2 + \epsilon)$ satisfies

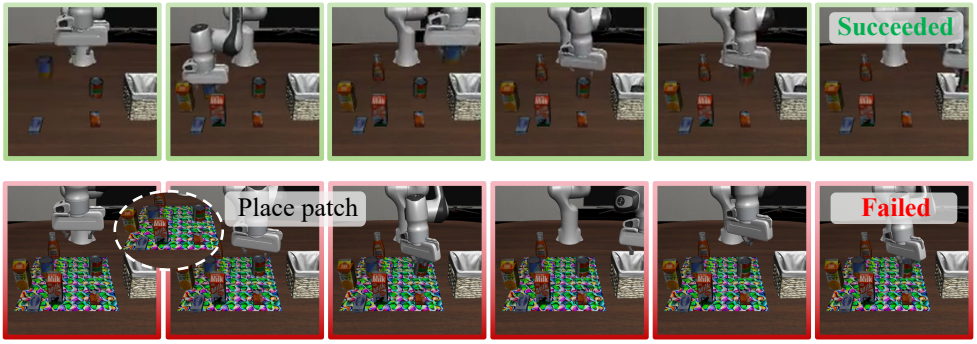
$$\left\| N_\epsilon(g) - \frac{g}{\|g\|_2} \right\|_2 = \frac{\epsilon}{\|g\|_2 + \epsilon}. \tag{53}$$

Consequently, if both hijacking gradients are nonzero and $\theta < \pi$, the normalized sum of their stabilized normalizations converges to u_{hij} as $\epsilon \to 0$. This approximation becomes ill-conditioned as the two gradients approach exact opposition.

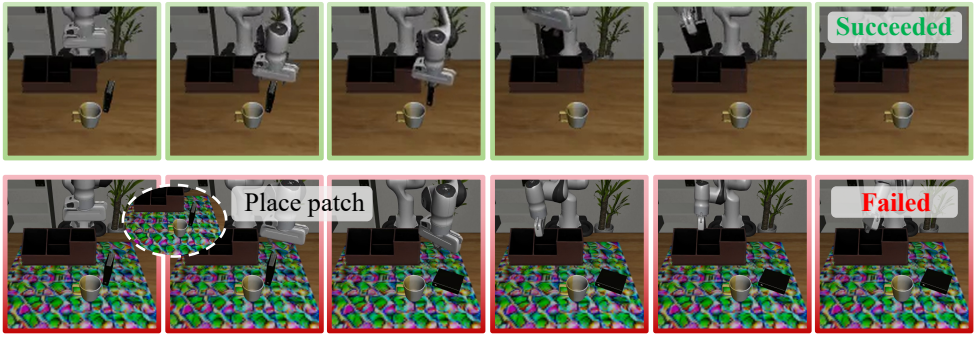
C VISUALIZATION OF SIMULATION RESULTS



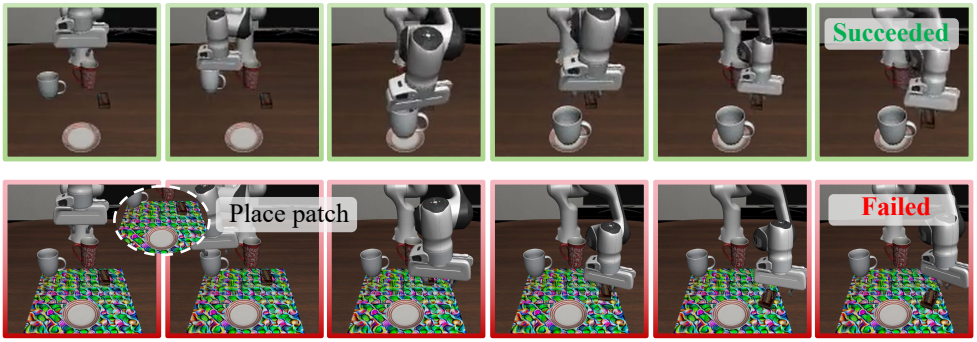
(a) Put the white mug on the left plate and put the yellow-and-white mug on the right plate.



(b) Put both the alphabet soup and the tomato sauce in the basket.



(c) Pick up the book and place it in the back compartment of the caddy.



(d) Put the white mug on the plate and put the chocolate pudding to the right of the plate.

Figure 6: In each subfigure, the green-bordered row shows the successful clean rollout, while the red-bordered row shows the failed rollout under the watermark attack. Frames are arranged chronologically from left to right. The dashed circle indicates the watermark patch location.