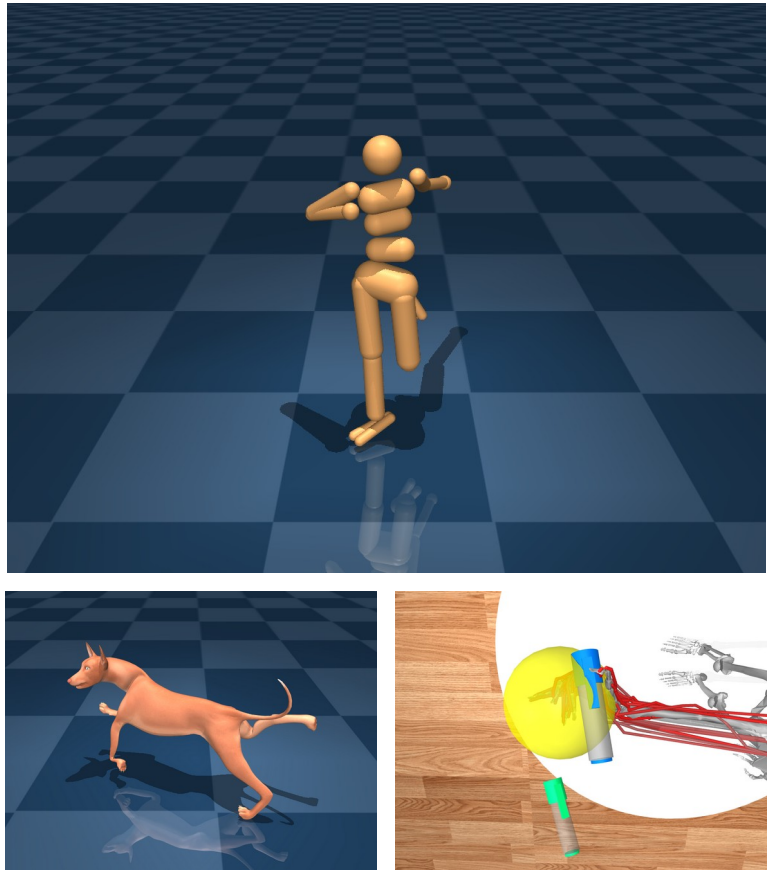
Stealthy World Model Manipulation via Data Poisoning

SWAAP

Can fine-tuning data be poisoned to reduce return without large data-level deviations?

Yibin Hu · Zizhan Zheng

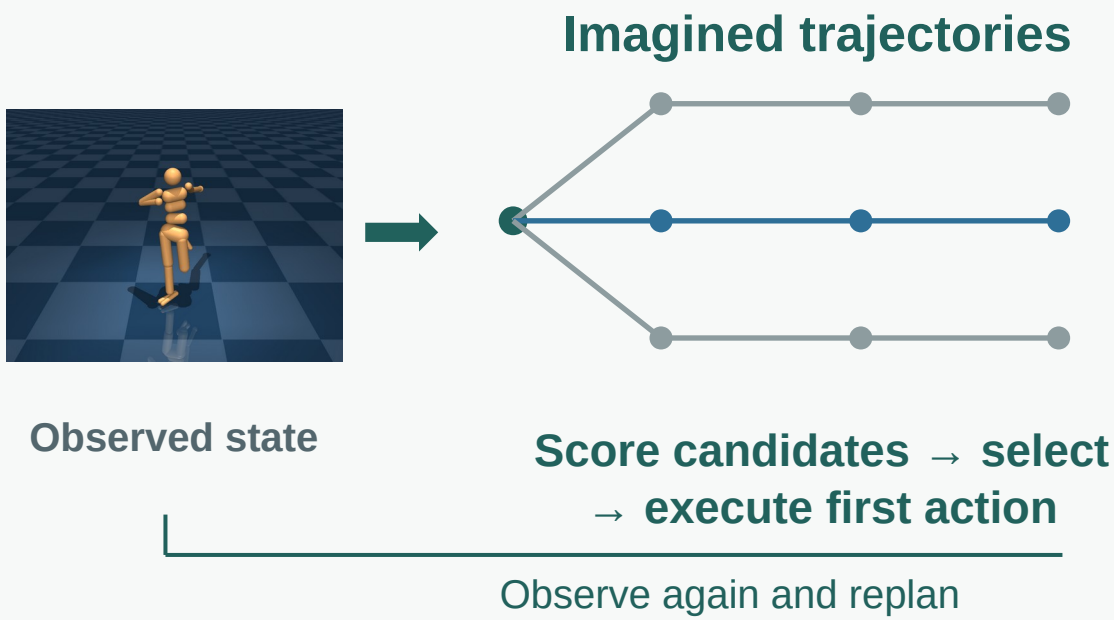
Department of Computer Science, Tulane University



Representative simulation tasks

A world model supports decisions, not just predictions

Model-predictive control (MPC)



Predict and optimize

Learned dynamics predict a successor state, often in latent space.

$$\hat{s}_{t+1} = P_\psi(\hat{s}_t, a_t)$$

The agent chooses a policy or plan that maximizes imagined discounted return.

$$J(P_\psi, \pi) = \mathbb{E}_{P_\psi, \pi}\left[\sum_{t=0}^{T-1} \gamma^t R_t\right]$$

A change in learned dynamics can change the planner's preferred actions.

Fine-tuning on local experience creates the attack surface

Pretraining provides a useful starting point, but local experience helps the model adapt to new dynamics or tasks.

Pretrained model
Starting dynamics

➡

Local transitions
Collected experience

➡

Fine-tune dynamics
Minimize prediction loss

➡

Plan with adapted model
Evaluate actual return

↑ **Poisoning entry point**

What the victim does

Fit dynamics on its recorded transitions.

$$\min_{\psi} \sum_{(s,a,s') \in D} \left\| s' - P_\psi(s,a) \right\|_2^2$$

What the attacker can change

$$(s_i, a_i, s_i') \longrightarrow (s_i, a_i, \tilde{s}_i')$$

White-box: knows the model and loss. Modify 10% of next-state targets; current states, actions and rewards unchanged.

The victim fine-tunes normally. Only the adaptation data are manipulated.

The attack must act through two coupled responses

Bilevel objective: reduce true-environment return while limiting model deviation

$$\min_{\widetilde{D}} \; J(P, \pi_{\theta^\star(\psi)}) + \lambda L(P_\psi, P)$$

$$P_\psi = \mathcal{F}(P_{\psi_0}, \widetilde{D}), \qquad \theta^\star(\psi) \in \arg\max_\theta J(P_\psi, \pi_\theta)$$

P: true dynamics • Pψ: adapted model • F: victim fine-tuning • Edit at most rp|D| successors

Stage 1 | Model space

Identify nearby dynamics that induce poor real-environment behavior.

➡

Stage 2 | Data space

Construct bounded data changes that move fine-tuning toward that target.

A harmful target model is useful only if its effect can be realized through data.

Camera-ready §2.3, Eq. (1), and §3

Identify dynamics that induce harmful planning

Model-space bilevel problem

$$\hat{\psi} \in \arg\min_{\psi} \left[J(P, \pi_{\theta^\star(\psi)}) + \lambda L(P_\psi, P) \right]$$

$$\theta^\star(\psi) \in \arg\max_{\theta} J(P_\psi, \pi_\theta)$$

Inner: agent response

Maximize imagined return under candidate dynamics.

Outer: harmfulness + model deviation

Reduce true-environment return; penalize departure from true dynamics.

Transition-gradient identity | fixed-policy return sensitivity

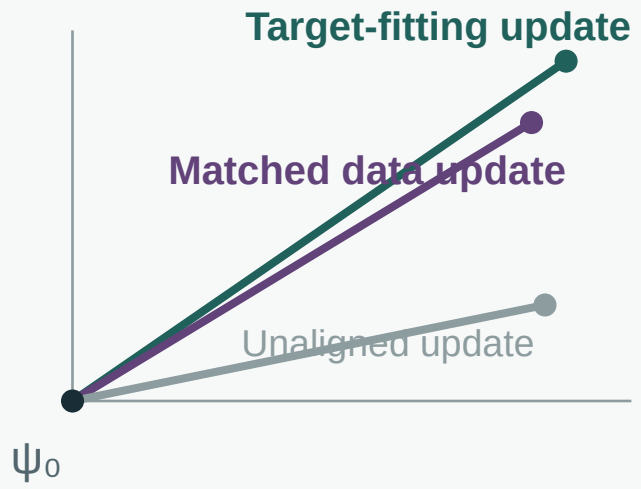
$$\nabla_\psi J(P_\psi, \pi_\theta) = \mathbb{E}_{P_\psi, \pi_\theta} \left[\sum_{t=0}^{T-1} \gamma^t \left(R_t + \gamma V_{t+1}(s_{t+1}) \right) \nabla_\psi \log P_\psi(s_{t+1} \mid s_t, a_t) \right]$$

Finite horizon, differentiable stochastic transitions; deterministic specialization in Appendix D.1.

The derivative supports a first-order BOME-style approximation of the bilevel response.

Realize the target at the clean residual scale

Match the training direction



Stealth-constrained gradient matching

Align the poisoned-dataset gradient with the target-fitting gradient at the pretrained weights.

$$L_P = (1-\alpha)\left[1-\cos(G_{\text{real}}, G_{\text{target}})\right] + \alpha\mathcal{R}_{\text{stealth}}$$

Match the corresponding clean prediction residuals:

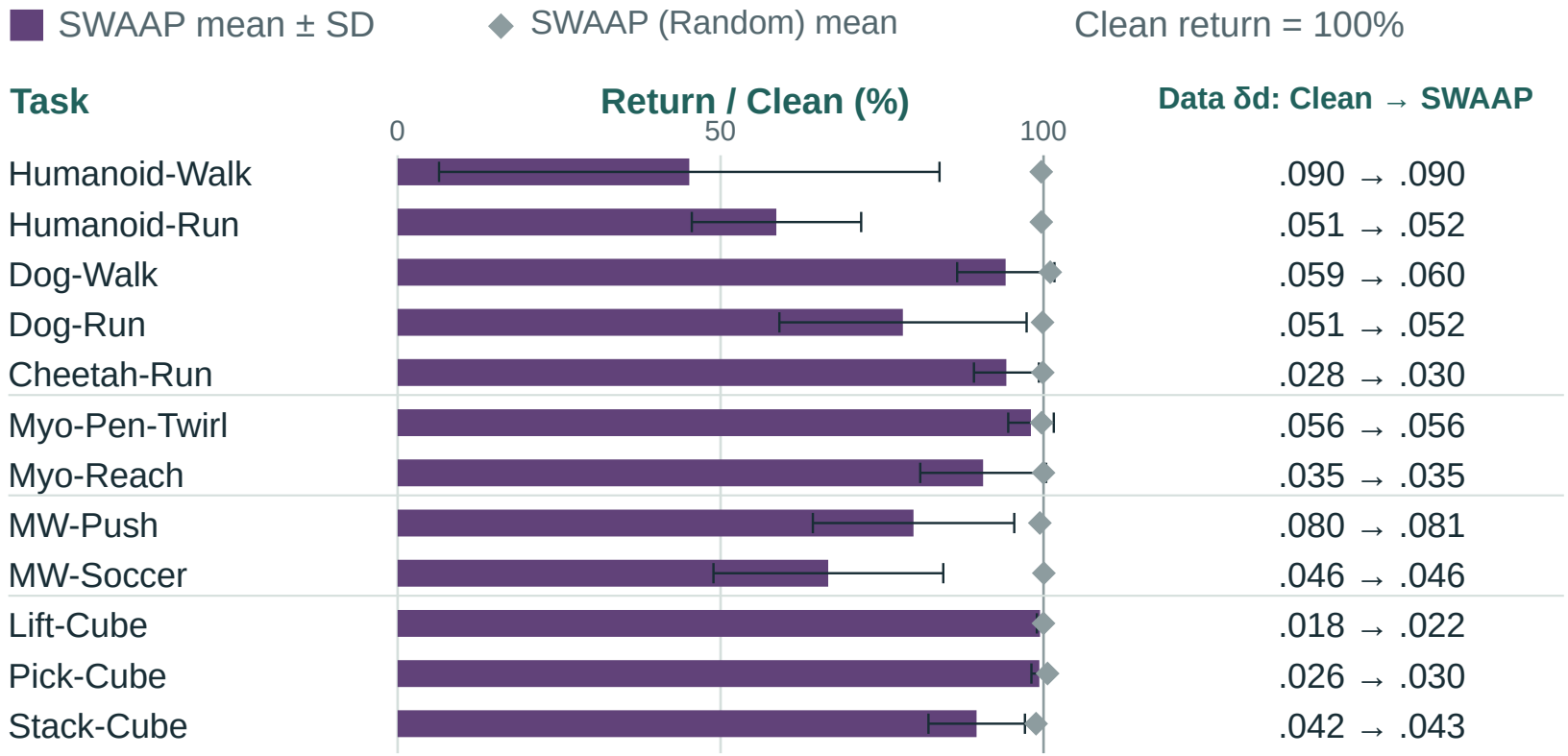
$$\mathcal{R}_{\text{stealth}} = \frac{1}{|S|}\sum_{i\in S}(\|\tilde{s}'_i - p_i\|_2 - \|s'_i - p_i\|_2)^2$$

Greal: loss gradient on the entire poisoned dataset; Gtarget: target-model fitting on clean reference inputs. S: selected indices; pi: pretrained prediction.

α controls the trade-off between gradient alignment and residual matching.

Large return losses with small data-level deviations

TD-MPC2 · 12 tasks · 500 of 5,000 successor targets edited



α = .9 except Lift .5, Pick/Stack .7. Three Stage 1 runs (Lift: two); three fits × 100 test episodes each. δd averages all adaptation rows.

Humanoid-Walk

54.7%

mean paired return reduction

0.090 → 0.090

normalized data error
Clean → SWAAP (rounded)

Effect varies by task.

Can the defender detect suspicious adaptation data?

Residual screening

Compare fine-tuning next-state targets with pretrained predictions before fine-tuning.

$$\delta_d = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{\|\tilde{s}'_i - p_i\|_2}{\|p_i\|_2}, \qquad p_i = P_{\psi_0}(s_i, a_i)$$

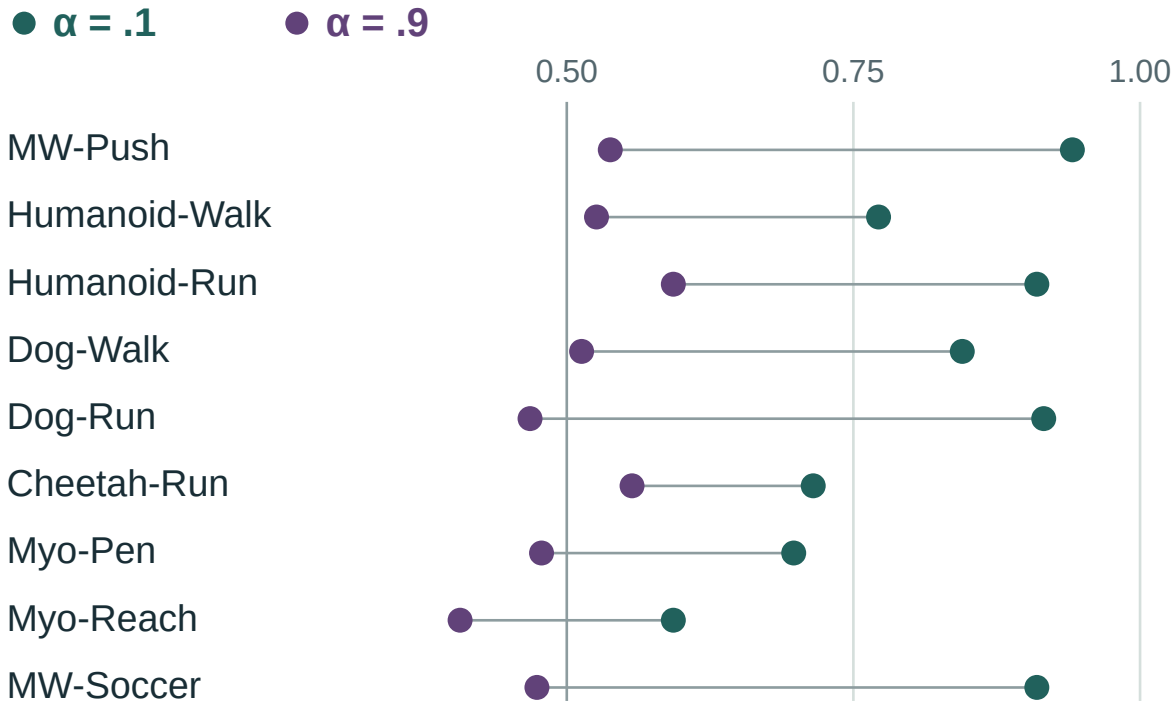
Residual-CUSUM

Scores sequences of unnormalized prediction residuals; thresholds calibrated on clean data.

0.506

Mean AUC at α = .9
Chance level: 0.50

Full ROC-AUC | lower separation at higher α



At α = .9, nominal-5% alarms match those of the unmodified source datasets. Detector-specific, not universal stealth.

TRIM and rollout monitoring address different risks

During fine-tuning | TRIM

Alternate dynamics fitting and retention of the lowest-loss 80% of transitions.

Task	Clean → TRIM	SWAAP → TRIM
Humanoid-Walk	857 → 858	387 → 484
Dog-Run	648 → 646	508 → 613
Myo-Reach	749 → 619	679 → 633

Partial recovery, remaining gaps, or clean-utility loss.

Equal epochs, not equal optimizer updates.

Test-time | rollout prediction error

Measure normalized prediction error on each model's executed trajectories.

$$\delta_m = \frac{1}{T}\sum_{t=0}^{T-1}\frac{\|P_\psi(s_t, a_t) - s'_t\|_2}{\|s'_t\|_2}$$

Humanoid-Walk .093 → .163

MW-Soccer .042 → .114

Reference: pretrained model on Clean rollouts. Larger errors can occur after poisoned fine-tuning.

Small data-level deviations do not imply small rollout prediction errors.

Does the vulnerability extend beyond one backbone?

DINO-WM | Visual planning

Push-T; history-dependent predictor; 10% of next-state targets modified.

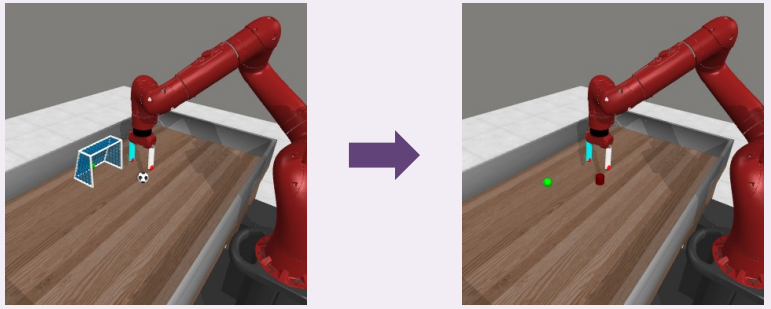
83.33% → 76.67%

Clean → SWAAP success, α = .9

6.67 ± 2.89 percentage-point decrease; same CUSUM alarms as unmodified data.

Three fits of one fixed target and dataset; 20 paired test goals.

Gray-box | cross-encoder transfer



Single-task Soccer surrogate → 80-task victim

Soccer: 5–9% lower return
Push: 11–18% lower return

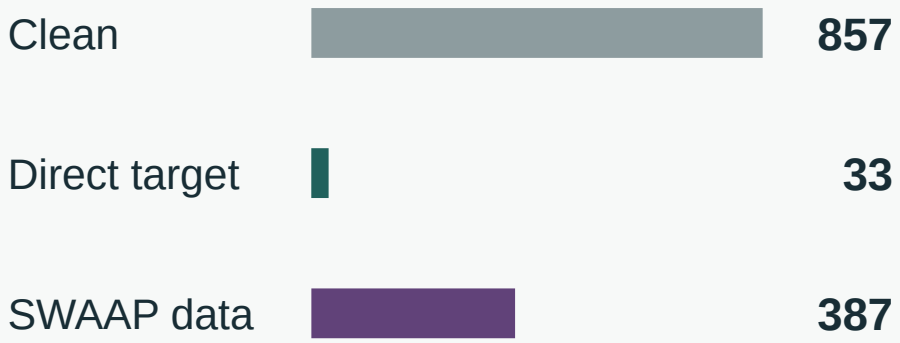
Selective transfer: median effect near zero across 80 tasks; some tasks improve.

Evidence across representations, but not uniform degradation or real-world validation.

Protect adaptation data and downstream planning

The target-realization gap

Humanoid-Walk mean return



Direct replacement has stronger access. Target harmfulness is not the same as data realizability.

What the results establish

1 Adaptation is a security-relevant step.

A bounded data intervention can change a later planner through normal fitting.

2 Stealth must be measured at multiple stages.

Data-level deviation, robust training and rollout prediction error answer different questions.

3 Realizability is an open problem.

Jointly selecting harmful and learnable targets remains future work.

github.com/yibhu/SWAAP

Scope: controlled simulation, mostly white-box, untargeted degradation.