Steering Robustness into World Action Models via Mechanistic Interpretability and Optimal Control

Jihoon Hong* Julian Skifstad* Qiyue Dai Alice Chan Glen Chou
Georgia Institute of Technology | *Equal Contribution
Robot World Models Workshop @ RSS 2026 · Sydney, Australia · July 17, 2026

Trustworthy Robotics Lab

Georgia Institute of Technology

Introduction & Setup

- WAMs jointly predict future states & actions, but remain brittle to camera, gripper, and noise perturbations.
- **Activation steering** is a powerful approach to influence model behavior at inference time by directly modifying activations during generation.
- We seek to **improve robustness** by steering **features related to environmental perturbations**.

Mechanistic Analysis of WAMs

- Project contrastive activations onto their top-3 PCs and fit a linear SVM; **hinge loss** measures separability (0=separable, 1=random).
- **Key result: separability is architecture-dependent**: strong in Cosmos-Policy/DiT4DiT, weak in LingBot-VA.

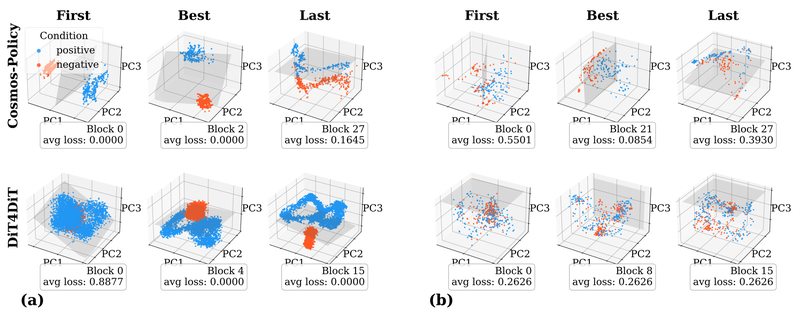

Fig. 2 — Cosmos-Policy & DiT4DiT activations separate cleanly under (a) noise and (b) camera perturbations.

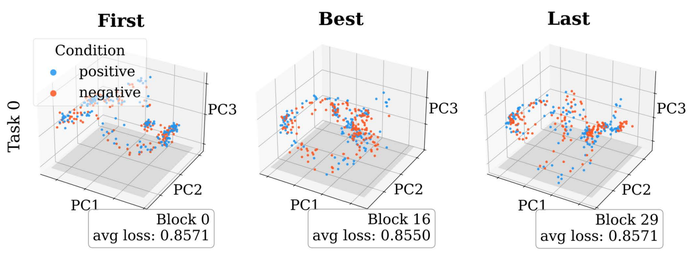

Fig. 4 — LingBot-VA activations overlap heavily — much weaker separability.

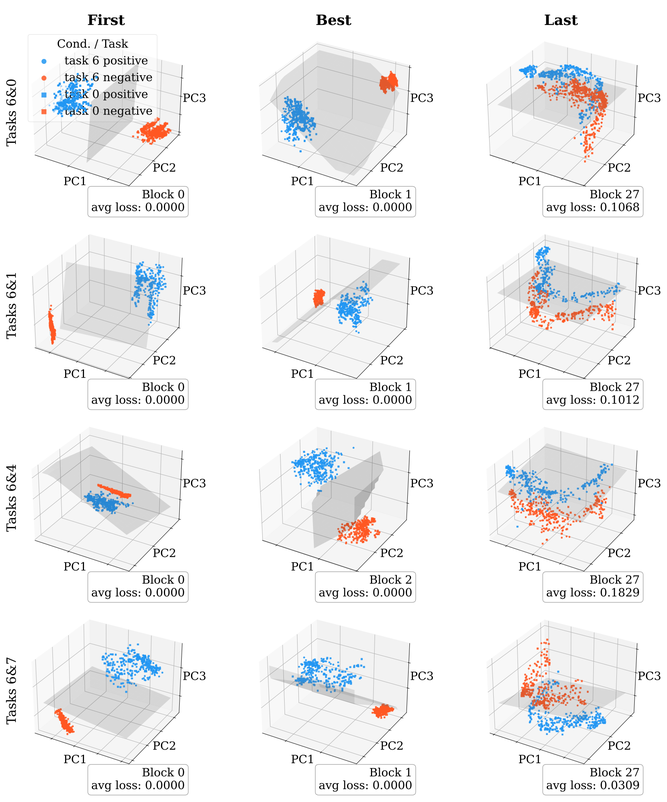

Fig. 3 — Some task pairs share separable directions, enabling cross-task steering.

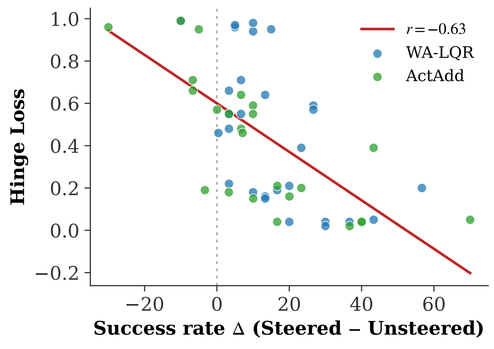

Fig. 7 — Lower hinge loss (better separability) predicts larger steering gains.

Key Takeaway

World Action Models can be steered to be more robust —**but only when contrastive activations are linearly separable.**

We propose an **inference-time alignment** approach that

- *is training-free and weight-preserving*
- *adapts online (WA-LQR), minimizing invasive edits*
- *improves Cosmos-Policy & DiT4DiT, but not LingBot-VA — exactly as predicted*

Separability $\underset{\textit{predicts}}{\Rightarrow}$ Steerability, but separability is architecture dependent

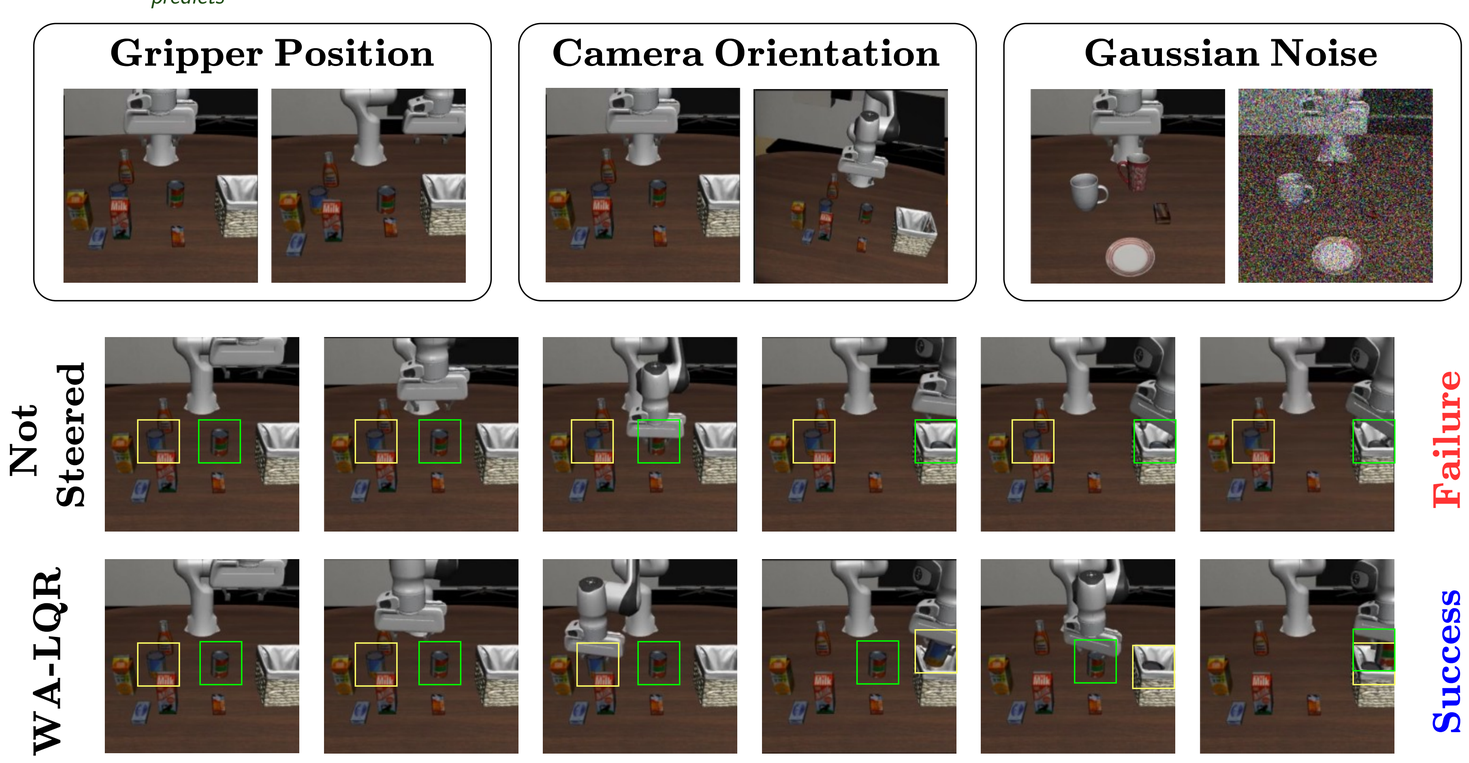

Fig. 1 — Overview: WA-LQR steers WAM activations toward robust behavior under camera, gripper, and noise perturbations.

Method 1: Open-Loop Steering (ActAdd)

Average contrastive directions into a steering vector, then add it at fixed strength γ:

$$a_{l,t} = \frac{1}{NH_a}\sum_{n,\tau} d^{(n)}_{l,t,\tau}, \qquad x_{l,t,\tau} \leftarrow x_{l,t,\tau} + \gamma\, a_{l,t}.$$

Simple and training-free, but *open-loop*: it ignores the current activation state and can oversteer.

Method 2: Closed-Loop Steering (WA-LQR)

1. **Project** into a low-dim contrastive subspace via SVD ($d_z \ll d_x$):
$$z_{l,t,\tau} = P_{l,t}\, x_{l,t,\tau} \in \mathbb{R}^{d_z}.$$
2. **Linearize latent dynamics** across transformer blocks:
$$\delta z_{l+1,t,\tau} \approx \widetilde{A}_{l,t}\,\delta z_{l,t,\tau} + \widetilde{B}_{l,t}\,\delta u_{l,t,\tau}.$$
3. **Track a feature setpoint** $\beta^{z,*}_{l,t} = \lambda\|e^z_{l,t}\|_2$ with online error $\alpha_{l,t,\tau} = \beta^{z,*}_{l,t} - (v^z_{l,t})^\top z_{l,t,\tau}$.
4. **Solve the LQR tracking problem** over the L transformer blocks (Riccati recursion, precomputed offline):
$$\min_{\{\delta u_{l,t}\}} \sum_l \left(\bar{\delta z}^\top_{l,t} Q_{l,t} \bar{\delta z}_{l,t} + \delta u^\top_{l,t} R^{\text{chunk}}_{l,t} \delta u_{l,t}\right) \text{ s.t. } \bar{\delta z}_{l+1,t} = \widetilde{A}_{l,t}\bar{\delta z}_{l,t} + \widetilde{B}_{l,t}\delta u_{l,t}.$$
5. **Closed-form control law** — apply the precomputed gains $K_{l,t}$ online:
$$\boxed{u^*_{l,t,\tau} = \bar{u}_{l,t,\tau} + K_{l,t,\tau}\alpha_{l,t,\tau}}$$

Intervention strength scales automatically with the *online tracking error* — steering only when, and as much as, needed. $R^{\text{chunk}}_{l,t}$ decays across the action chunk, so early timesteps are steered most.

Why Is WA-LQR Tractable?

Two empirical properties of WAM activations justify the reduced-order LQR design above.

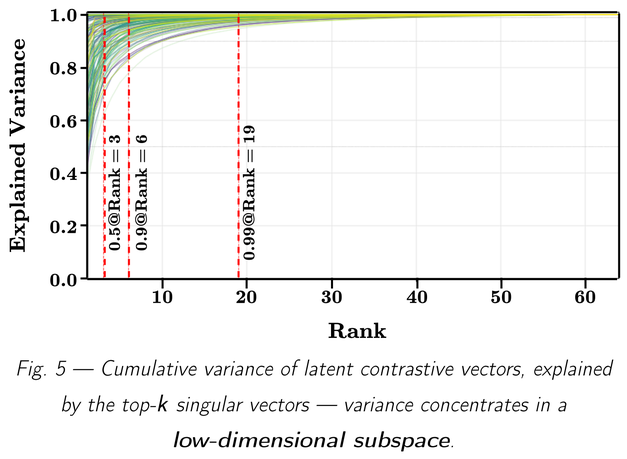

*Fig. 5 — Cumulative variance of latent contrastive vectors, explained by the top-**k** singular vectors — variance concentrates in a **low-dimensional subspace**.*

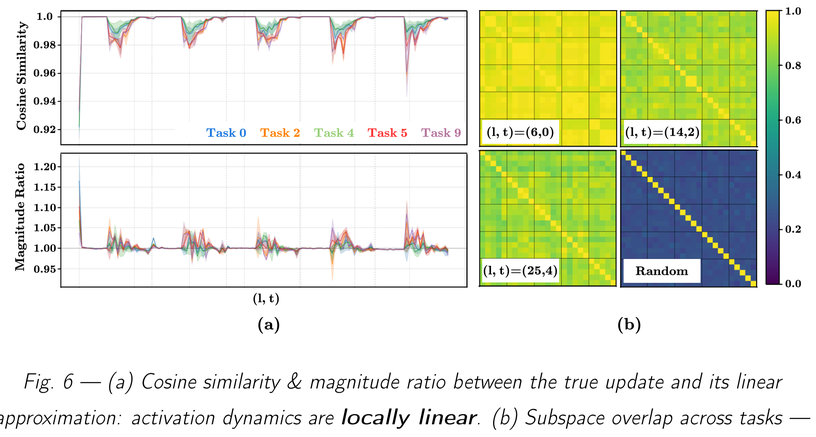

*Fig. 6 — (a) Cosine similarity & magnitude ratio between the true update and its linear approximation: activation dynamics are **locally linear**. (b) Subspace overlap across tasks — one controller **transfers**.*

Results

Table 1. Success rate on LIBERO-10 under perturbations.

Model	Perturbation	No Steer	ActAdd	WA-LQR
Cosmos-Policy	Camera Orientation	46.0%	49.3%	**59.3%**
	Init. Gripper Pos.	61.3%	63.3%	**72.7%**
	Gaussian Noise	26.7%	**67.3%**	58.7%
DiT4DiT	Init. Gripper Pos.	65.7%	65.8%	**71.7%**
	Gaussian Noise	15.6%	43.3%	**48.9%**
LingBot-VA	Camera Orientation	45.0%	43.3%	48.3%

- WA-LQR improves Cosmos-Policy & DiT4DiT robustness across perturbation types.
- LingBot-VA gains are small/inconsistent — as predicted by its weak separability.

Summary & Limitations

- We study **mechanistic interpretability** and **activation steering** for robustifying World Action Models (WAMs).
- Linear **separability** of robustness-relevant features **predicts steerability** — and both are architecture-dependent.
- **WA-LQR**, exploits locally-linear activation dynamics for training-free robustness gains.
- No method yet predicts *a priori* which tasks share transferable directions.
- Steerability depends strongly on the backbone architecture.