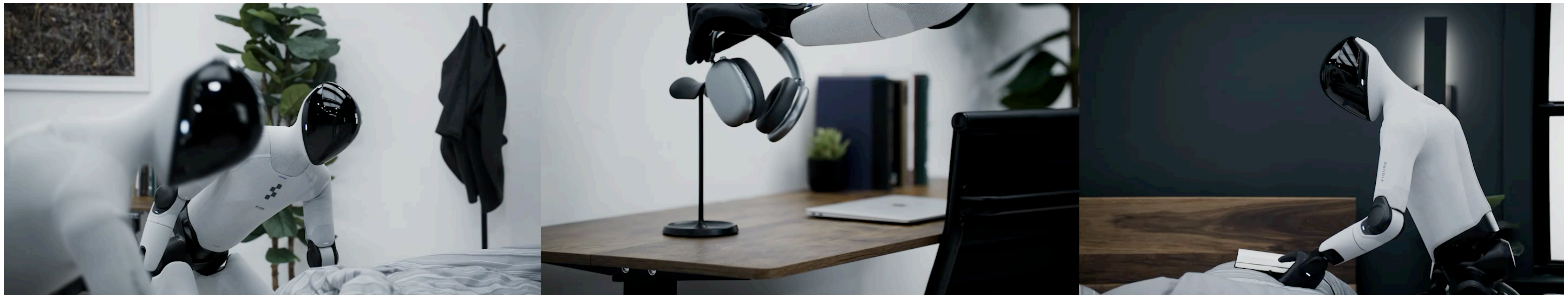
Latent Representations for Control Design

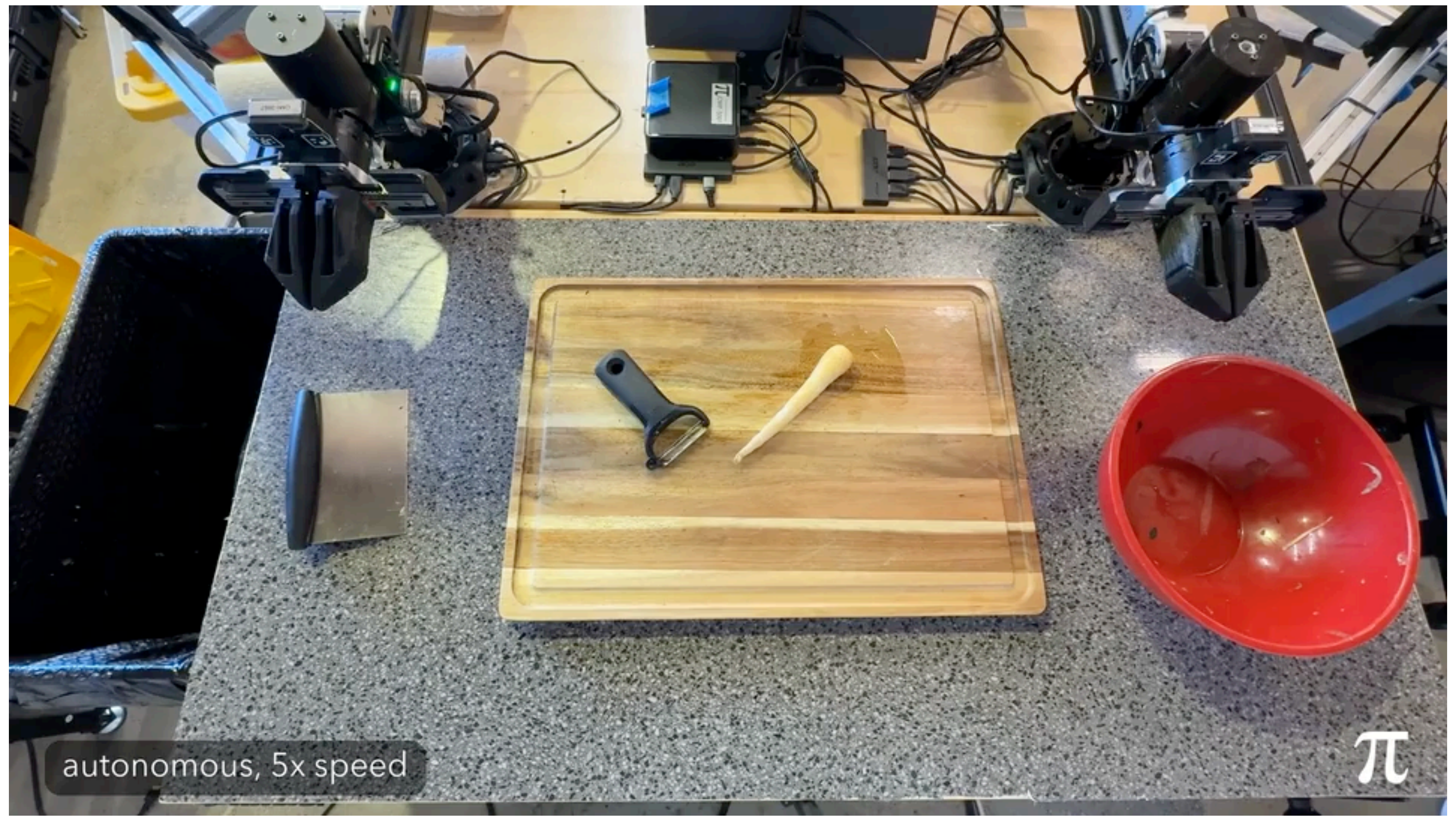
Stephen Tu
University of Southern California
May 13, 2026

USC Viterbi
School of Engineering
Ming Hsieh Department of Electrical and Computer Engineering

Meanwhile in Robot Land…



[Helix-02, Figure AI]

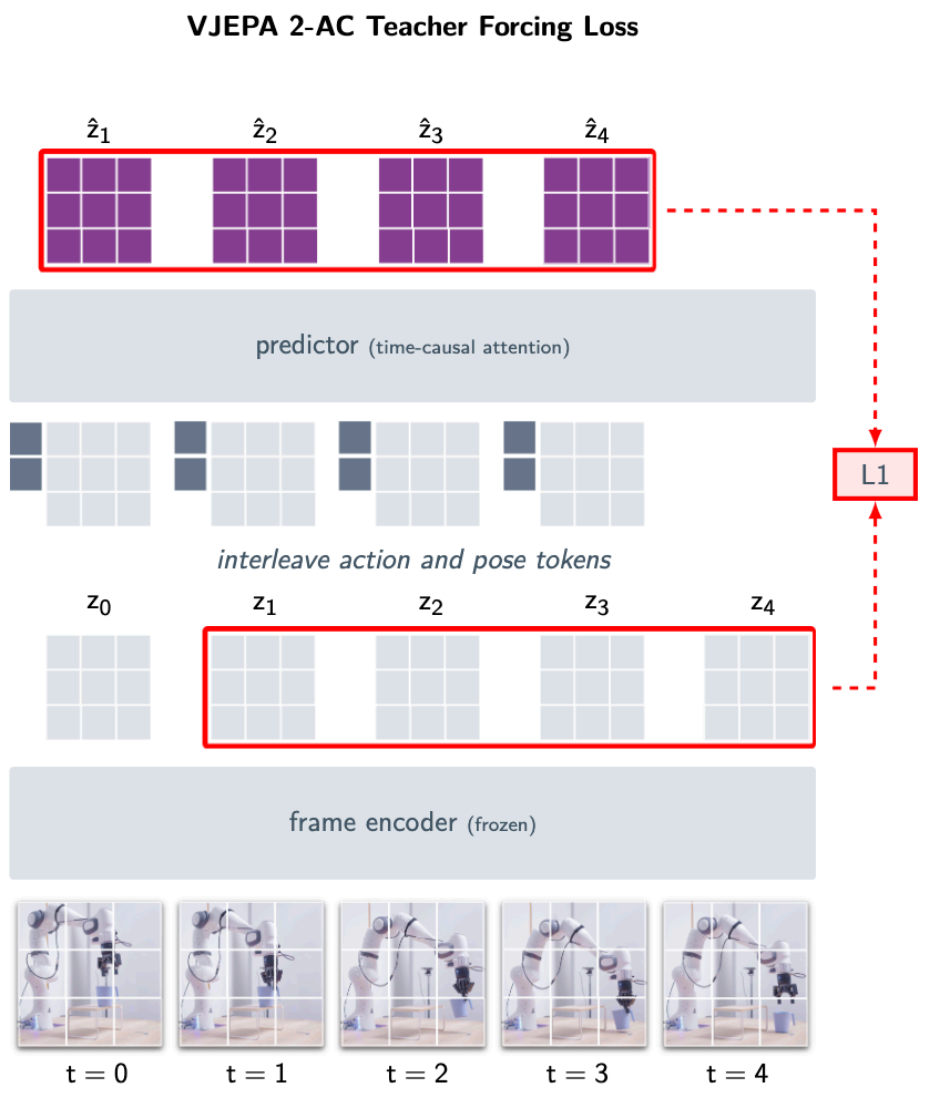


[Physical Intelligence, $\pi_{0.7}$]

[Waymo]

Imitation vs. World Models

- Most SoTA robot demonstrations are trained via **imitation learning.** [Unitree D1-T]
    - **Data bottleneck—** how to collect “LLM scale” human demos?
- **World models** provide an exciting alternative/augmentation:
    - Can be trained on diverse set of non-optimal data.
    - Can also be used for **analysis** (safety, reachability, etc.).



DINO-WM: World Models on Pre-trained Visual Features enable Zero-shot Planning

Gaoyue Zhou [1] Hengkai Pan [1] Yann LeCun [1 2] Lerrel Pinto [1]

Abstract

The ability to predict future outcomes given control actions is fundamental for physical reasoning. However, such predictive models, often called world models, remain challenging to learn and are typically developed for task-specific solutions with online policy learning. To unlock world models' true potential, we argue that they should 1)

2024; Hansen et al., 2024; Haldar et al., 2024; Jia et al., 2024). Despite this progress, generalization remains a major challenge (Zhou et al., 2023). Existing approaches predominantly rely on policies that, once trained, operate in a feed-forward manner during deployment—mapping observations to actions without any further optimization or reasoning. Under this framework, successful generalization inherently requires agents to possess solutions to all possible tasks and scenarios once training is complete, which is

NVIDIA 2025-7-11

Cosmos World Foundation Model Platform for Physical AI

NVIDIA[1]

Abstract

Physical AI needs to be trained digitally first. It needs a digital twin of itself, the policy model, and a digital twin of the world, the world model. In this paper, we present the Cosmos World Foundation Model Platform to help developers build customized world models for their Physical AI setups. We position a world foundation model as a general-purpose world model that can be fine-tuned into customized world models for downstream applications. Our platform covers a video curation pipeline, pre-trained world foundation models, examples of post-training of pre-trained world foundation models, and video tokenizers. To help Physical AI builders solve the most critical problems of our society, we make Cosmos open-source and our models open-weight with permissive licenses available via NVIDIA Cosmos-Predict1.

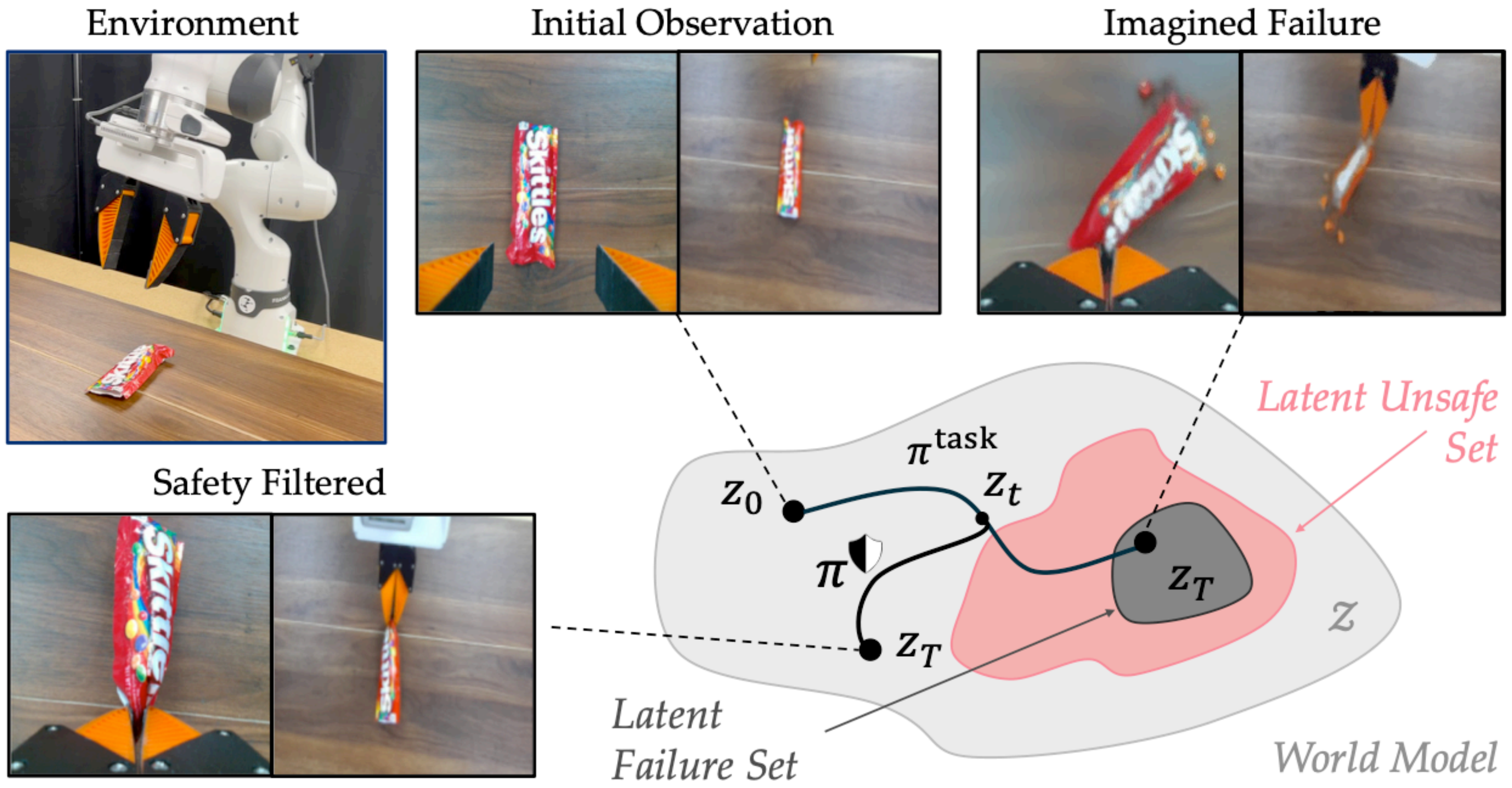
V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning

Mahmoud Assran[1,*], Adrien Bardes[1,*], David Fan[1,*], Quentin Garrido[1,*], Russell Howes[1,*], Mojtaba Komeili[1,*], Matthew Muckley[1,*], Ammar Rizvi[1,*], Claire Roberts[1,*], Koustuv Sinha[1,*], Artem Zholus[1,2,*], Sergio Arnaud[1,*], Abha Gejji[1,*], Ada Martin[1,*], Francois Robert Hogan[1,*], Daniel Dugas[1,*], Piotr Bojanowski[1], Vasil Khalidov[1], Patrick Labatut[1], Francisco Massa[1], Marc Szafraniec[1], Kapil Krishnakumar[1], Yong Li[1], Xiaodong Ma[1], Sarath Chandar[2], Franziska Meier[1,*], Yann LeCun[1,*], Michael Rabbat[1,*], Nicolas Ballas[1,*]

[1]FAIR at Meta, [2]Mila – Quebec AI Institute and Polytechnique Montréal
*Core Team

Reachability with (Latent) World Models

- Latent world models enable reachability analysis for **safety**:



Generalizing Safety Beyond Collision-Avoidance via Latent-Space Reachability Analysis

Kensuke Nakamura
Carnegie Mellon University
kensuken@andrew.cmu.edu

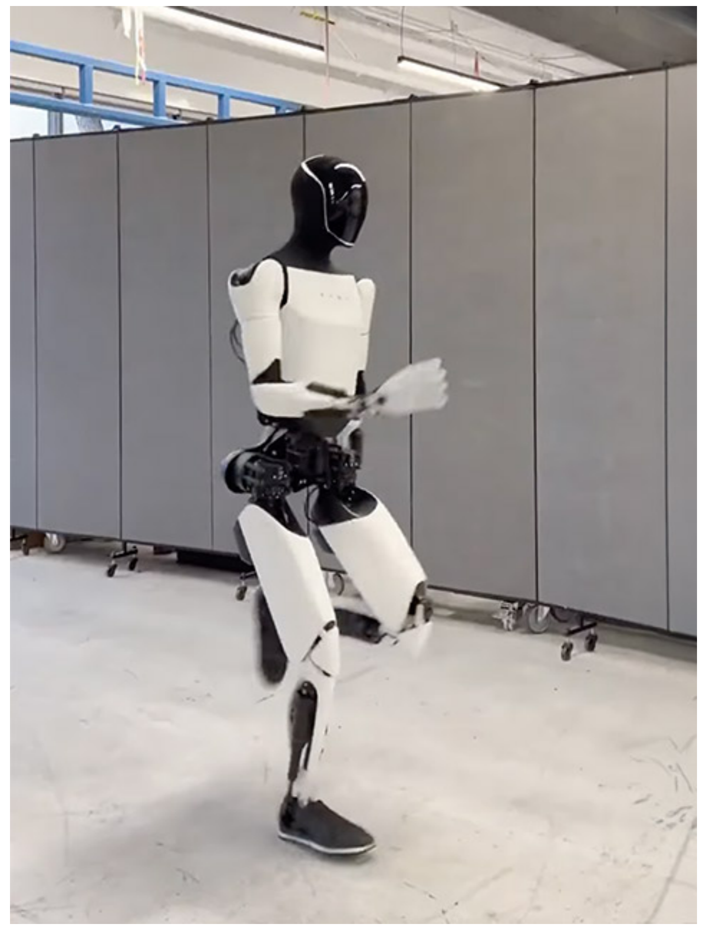
Lasse Peters
Delft University of Technology
l.peters@tudelft.nl

Andrea Bajcsy
Carnegie Mellon University
abajcsy@andrew.cmu.edu

Systems Analysis with Latent Models

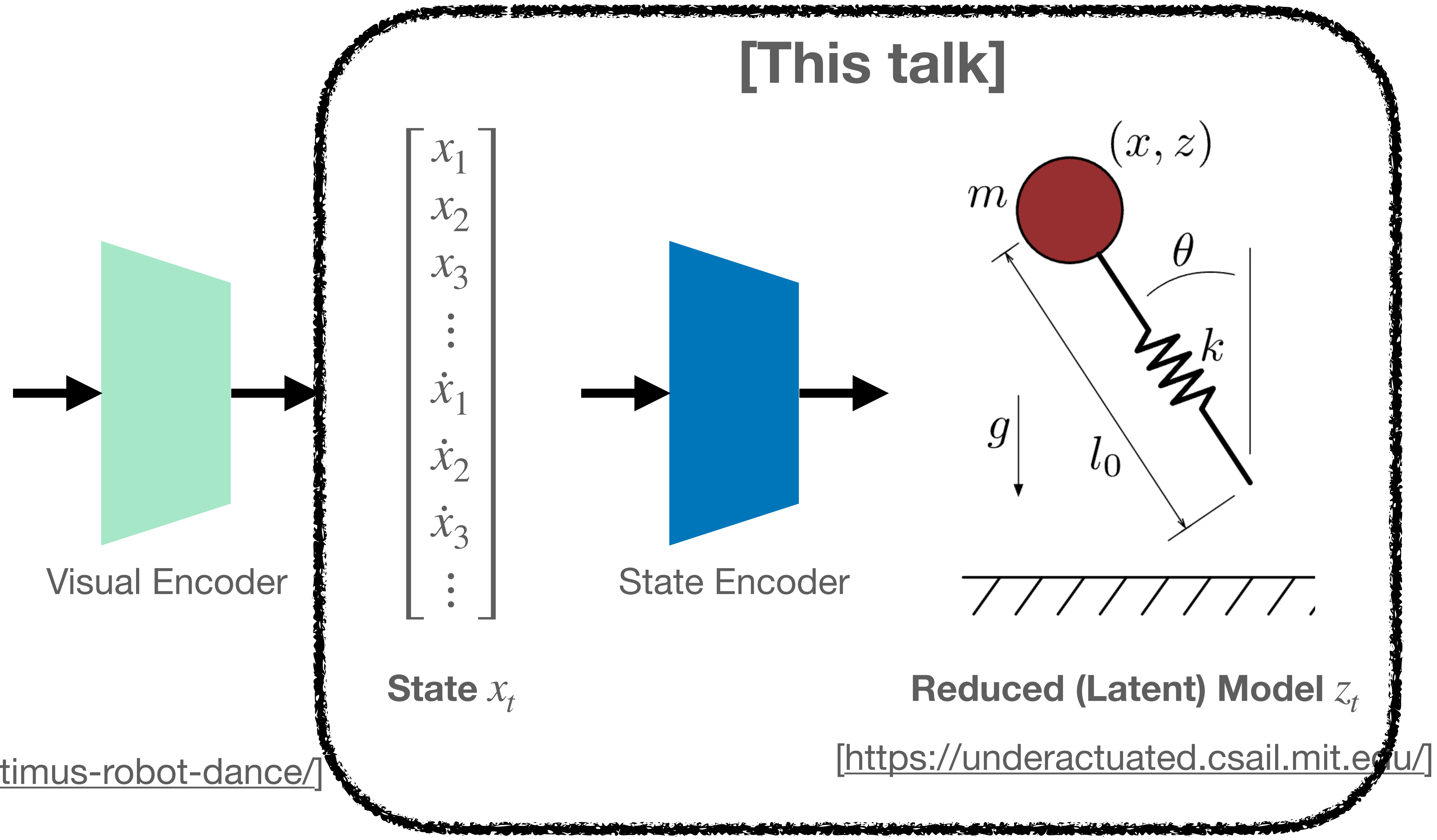
- What does certifying stability/reachability analysis/barrier functions on a **latent model** imply about closed-loop behavior on the **true system**?

- **Key sources of error in the pipeline:**



Observations y_t

[https://www.techeblog.com/tesla-optimus-robot-dance/]



[https://underactuated.csail.mit.edu/]

Latent Models for Control Design



Joint work with P. Lutkus, K. Wang, and L. Lindemann

Latent Models for Control Design

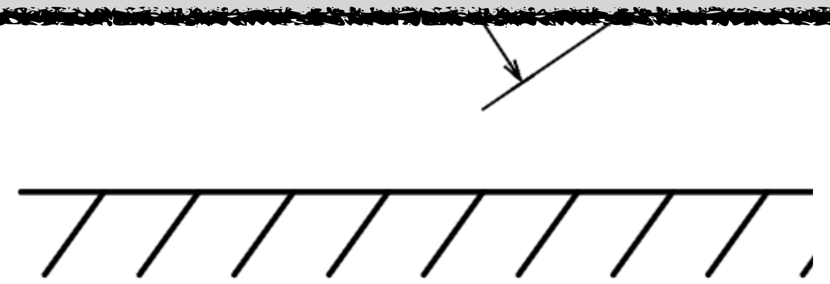
$\dot{x} = f(x, u)$

- **Goal:** Transfer analysis guarantees for closed-loop latent model onto the closed-loop original system.
- **Q1:** How do we measure model discrepancy between original and latent models?
- **Q2:** How do we analyze the original closed-loop system utilizing the latent system guarantees?

[https://www.techeblog.com/tesla-optimus-robot-dance/]

[~50 DoF + Environment]

$z\|)$



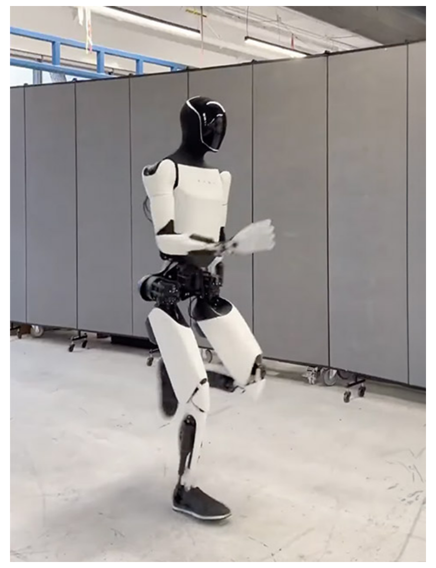
[More computationally efficient in lower dimension.]

[https://underactuated.csail.mit.edu/]

[2 DoF (SLIP Model)]

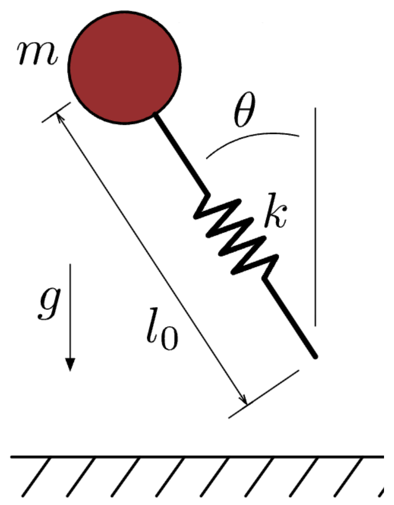
Problem Setup

$\dot{x} = f(x, u)$



[https://www.techeblog.com/tesla-optimus-robot-dance/]

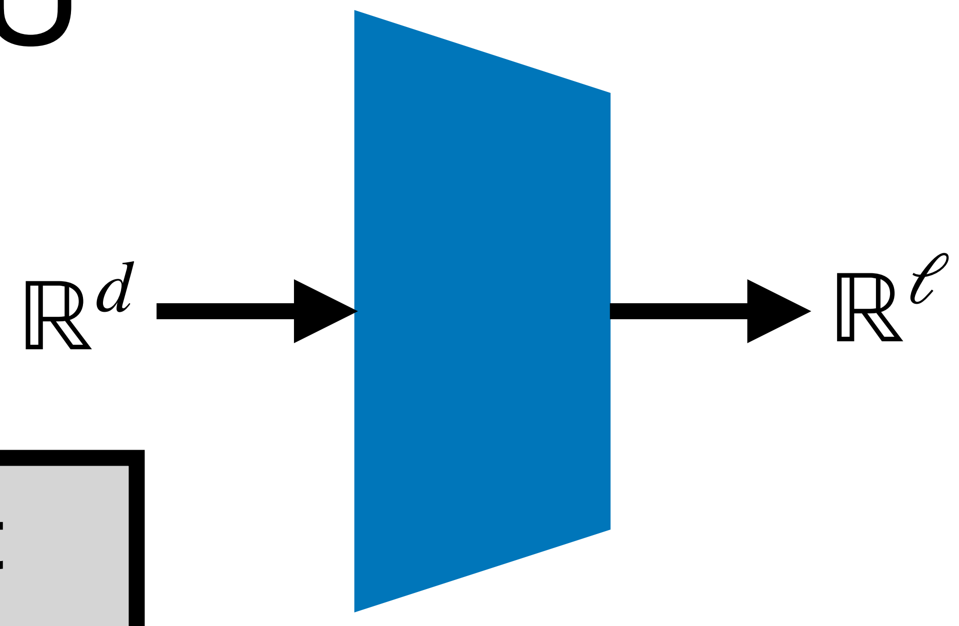
$\dot{z} = f_z(z, u)$



[https://underactuated.csail.mit.edu/]

- Original system: $\dot{x} = f(x, u)$ ($x \in \mathbb{R}^d$)
- Latent system + policy: $\dot{z} = f_z(z, u)$ ($z \in \mathbb{R}^\ell$), $\pi(z) : \mathbb{R}^\ell \mapsto$ U
- Encoder: $E(x) : \mathbb{R}^d \mapsto \mathbb{R}^\ell$
- Induced policy: $\bar{\pi}(x) := \pi(E(x))$
- **Goal**:

Special case of static output feedback with output $E(x)$.



$E(x)$

$\dot{z} = f_z^{(\pi)}(z) := f_z(z, \pi(z))$ satisfies property A $\Longrightarrow$

$\dot{x} = f^{(\bar{\pi})}(x) := f(x, \bar{\pi}(x))$ satisfies property A.

A = exponential Lyapunov stability for the rest of this talk.

Comparing Dynamics

- First attempt:

$$\|f^{(\bar{\pi})}(x) - f_z^{(\pi)}(E(x))\| = \| \quad - \quad \|$$

- Use Jacobian $\dfrac{\partial E}{\partial x}(x) \in \mathbb{R}^{\ell \times d}$ to adjust dimensions [Tabuada et al. 08, Szalai 23]:
  - $\left\| \dfrac{\partial E}{\partial x}(x) f^{(\bar{\pi})}(x) - f_z^{(\pi)}(E(x)) \right\| \leqslant \gamma$ **[Approximate Forward Conjugacy]**
  - $\left\| f^{(\bar{\pi})}(x) - \dfrac{\partial E}{\partial x}(x)^{\dagger} f_z^{(\pi)}(E(x)) \right\| \leqslant \gamma$ **[Approximate Backward Conjugacy]**

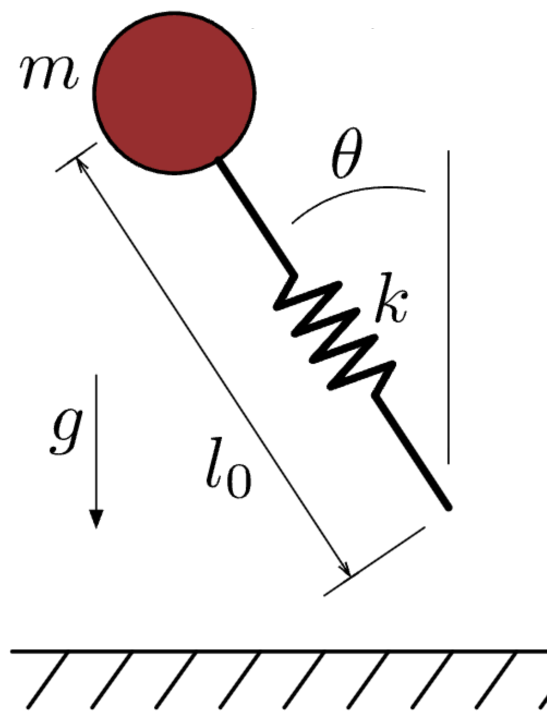
Latent Lyapunov Stability $\Rightarrow$ Original System Stability

- **Theorem** [LWLT25]**:** Suppose $V(z)$ is a Lyapunov function for the **latent** system:

$$\dot{V}(z) = \langle \nabla V(z), f_z^{(\pi)}(z) \rangle \leqslant -\rho V(z).$$

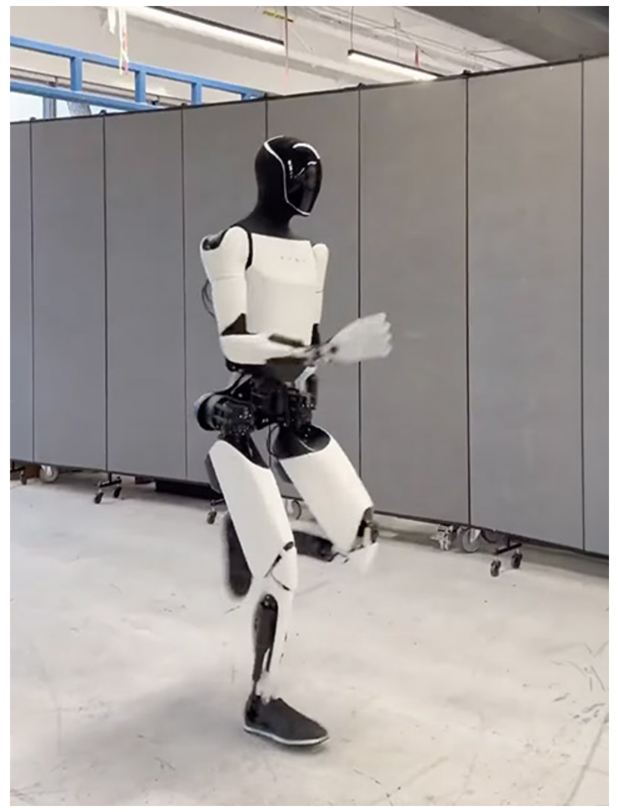
- The candidate Lyapunov function $\bar{V}(x) := V(E(x))$ for the **original system** satisfies:

$\dot{z} = f_z(z, u)$



$\dot{V} \leqslant -\rho V$

$\dot{x} = f(x, u)$



$\dot{\bar{V}} \leqslant -\rho \bar{V} + L\gamma$

$$\underbrace{\| \partial E(x) f^{(\bar{\pi})}(x) - f_z^{(\pi)}(E(x)) \|_\infty \leqslant \gamma}_{\text{Approx. forward conjugacy}}$$

OR

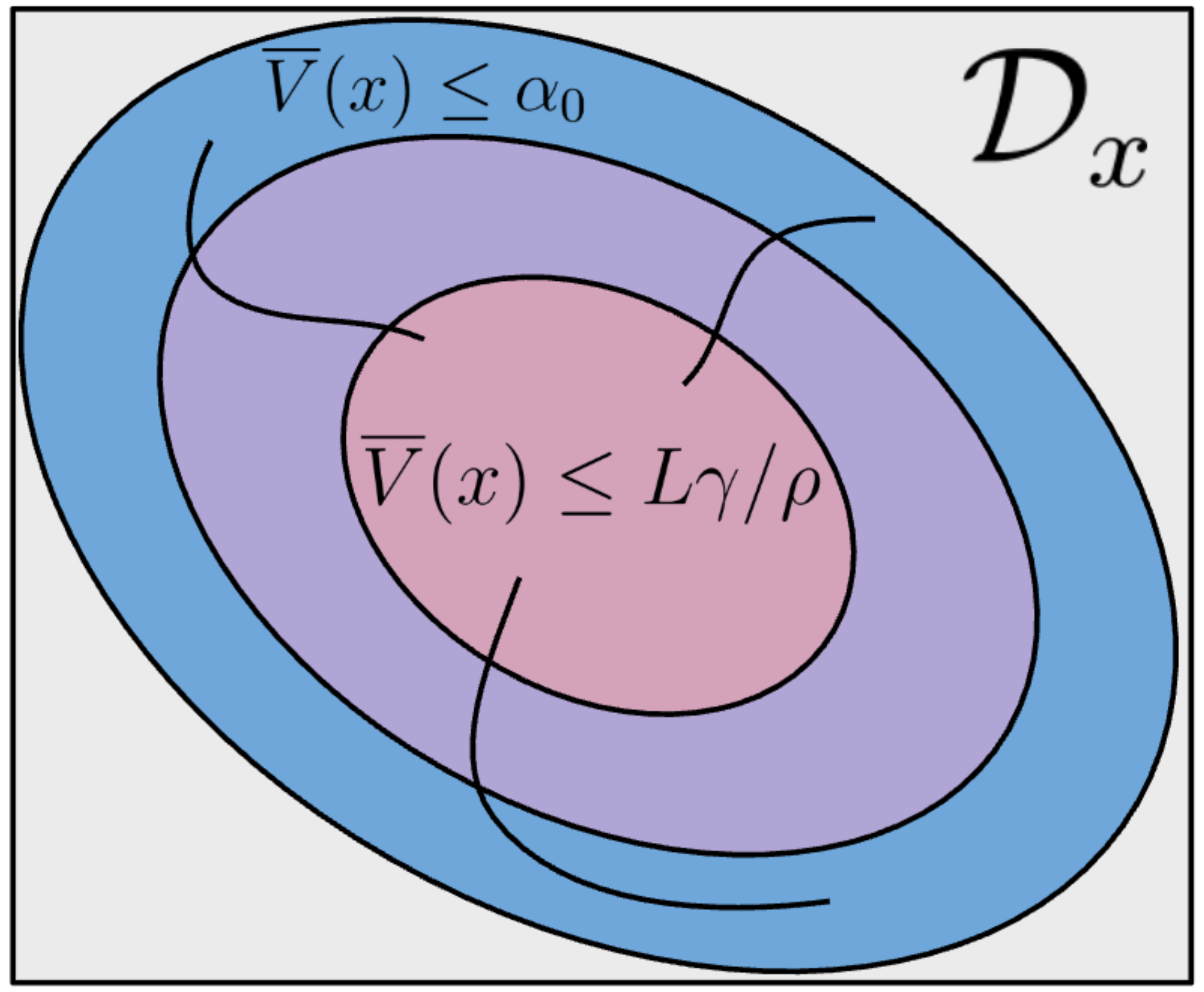
$$\underbrace{\| f^{(\bar{\pi})}(x) - \partial E^{\dagger}(x) f_z^{(\pi)}(E(x)) \|_\infty \leqslant \gamma}_{\text{Approx. backward conjugacy}}$$

Latent Lyapunov Stability $\Rightarrow$ Original System Stability

- What does $\dot{\bar{V}} \leqslant -\rho\bar{V} + L\gamma$ imply?

- The set $\{\bar{V}(x) \leqslant L\gamma/\rho\}$ is an **attractive set** — $x(t)$ flow towards this set.

- If $\gamma = 0$, $x(t)$ flows towards $\{\bar{V}(x) = 0\}$.

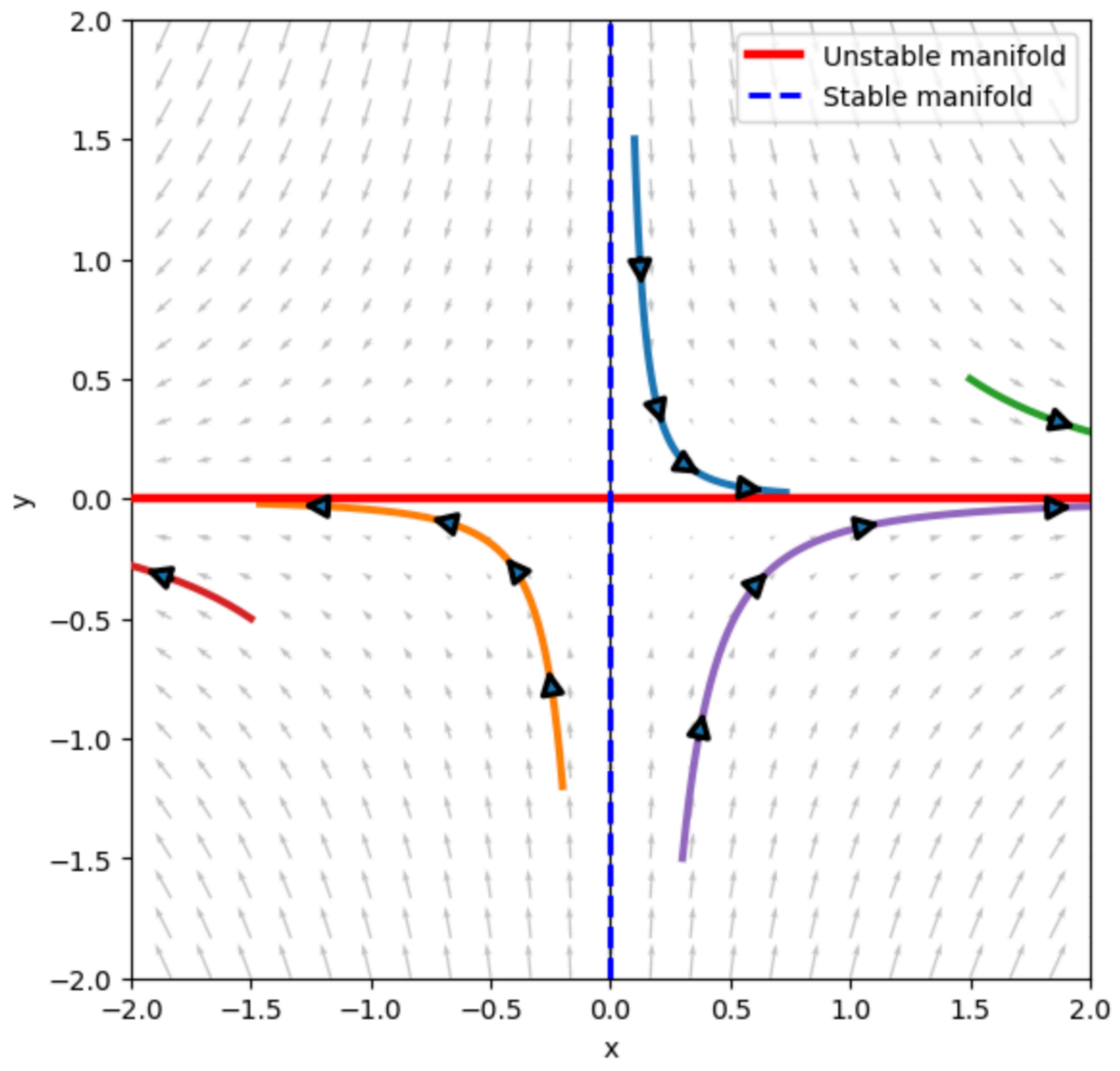


Zero-Manifold Convergence

- $x(t) \to \{\bar{V}(x) = 0\}$ implies $x(t) \to \{E(x) = 0\}$.

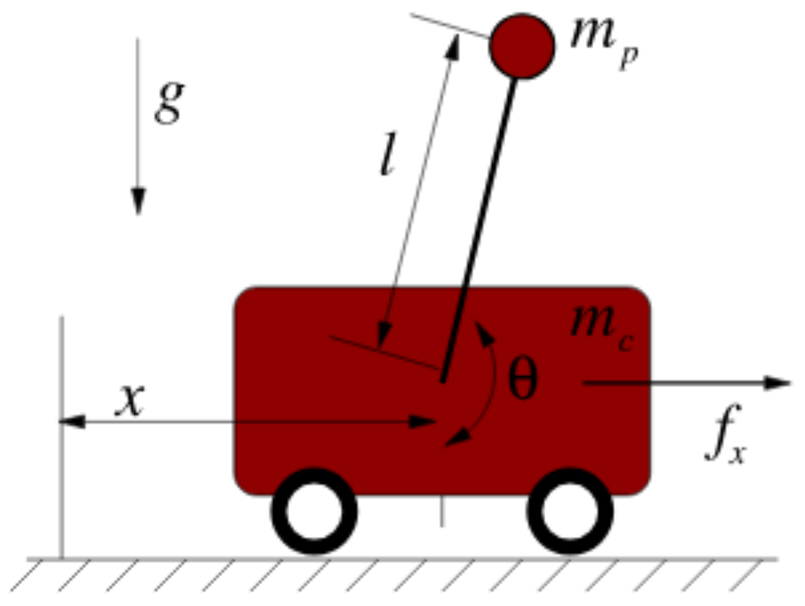
- The zero-surface $\mathscr{M} := \{x \in \mathbb{R}^d \mid E(x) = 0\}$ is generally a manifold of dimension $d - \ell$.

- The behavior of the system on $\mathscr{M}$ is not directly controllable — not guaranteed to be stable.



- **No free lunch:** In general, you have to give something up (similar to partial feedback linearization).

Case Study: Cart-pole



https://underactuated.csail.mit.edu/

- **Task**: Stabilize cart-pole using a latent LQR controller.
- Learn encoder $E : \mathbb{R}^4 \mapsto \mathbb{R}^3$, decoder $D : \mathbb{R}^3 \mapsto \mathbb{R}^4$, latent dynamics model $z_{t+1} = A(z_t)z_t + B(z_t)u_t$:
  - $J_{\text{bwd}}(E, D, \Phi_z) := T^{-1} \sum_{t=0}^{T-1} \mathbb{E}_{(x,\mathbf{u})} \|\Phi^t(x, \mathbf{u}) - D(\Phi_z^t(E(x), \mathbf{u}))\|^2$

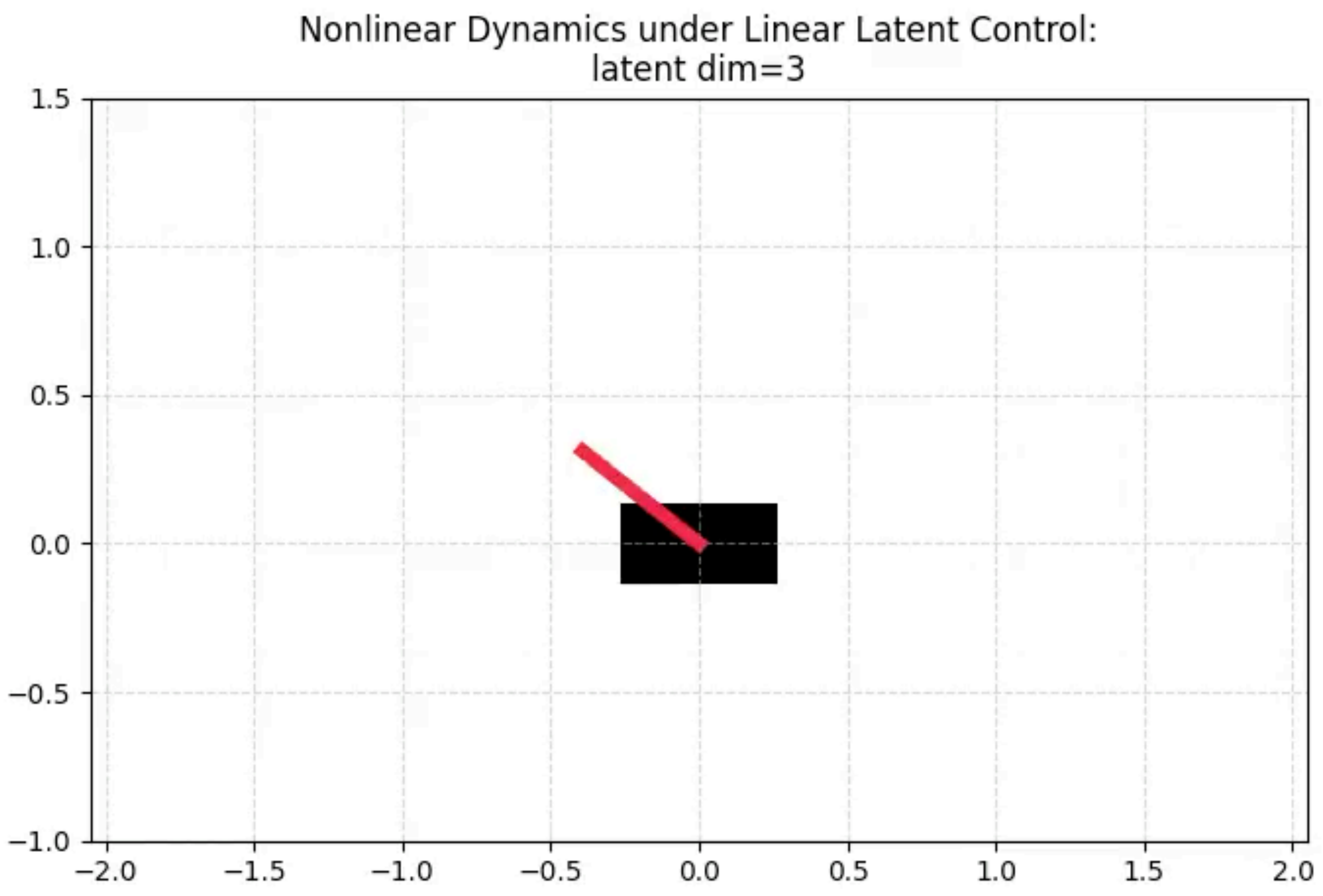
Note: Discrete-time losses.
(Neural-ODEs are difficult to train!)

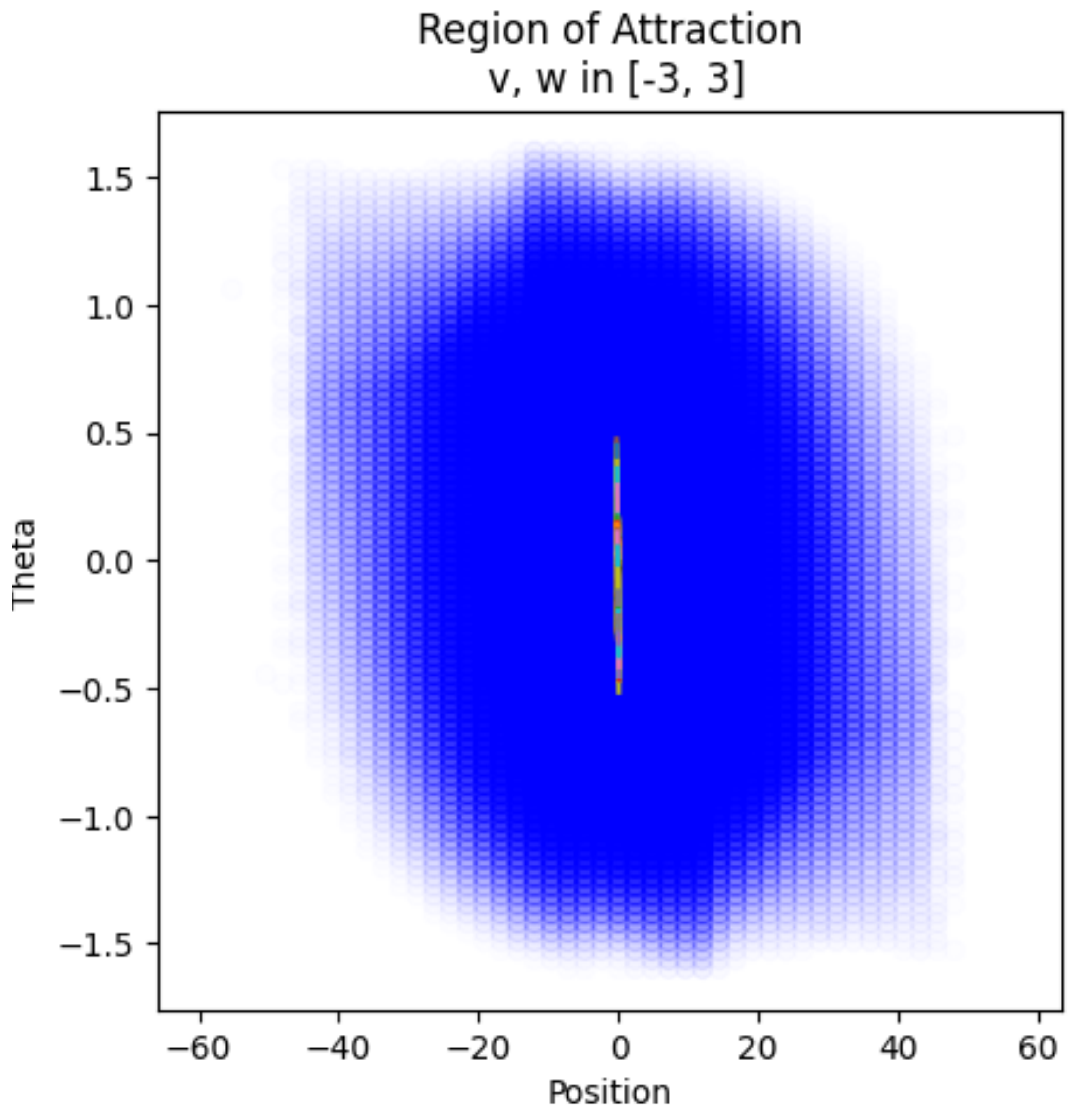
[Approx. Backward Conjugacy]

  - **Note:** Includes auto-encoder loss $\mathbb{E}_x \|x - D \circ E(x)\|^2$ as one term
- Closed-loop controller: $\pi(z) = \mathrm{LQR}\left(A(z_0), B(z_0)\right), \quad z_0 := E(0).$

Upright equilibrium

Case Study: Cart-pole





Learning Stable Zero Dynamics

Joint work with P. Lutkus and L. Lindemann

Learning Stable Zero Dynamics

- Previous theory is agnostic to **how** encoder/decoder + latent Lyapunov function is recovered.

- Our experiment uses a specific recipe: **multi-step backwards losses**

$$J_{\mathrm{bwd}}(E, D, \Phi_z) := \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_{(x,\mathbf{u})} \|\Phi^t(x, \mathbf{u}) - D(\Phi_z^t(E(x), \mathbf{u}))\|^2.$$

- **Q:** Does minimizing $J_{\mathrm{bwd}}(E, D, \Phi_z)$ **implicitly** yield stable zero dynamics?

Idealized Encoder for Linear Systems

- Let's use LTI systems and linear encoder/decoders in order to build our understanding:
  - $x_{t+1} = Ax_t + Bu_t, \quad A \in \mathbb{R}^{d\times d}, B \in \mathbb{R}^{d\times p},$
  - $z = Ex, \quad x = Dz, \quad E \in \mathbb{R}^{\ell\times d}, D \in \mathbb{R}^{d\times \ell}.$

- Define **stable** (σ_-) and **unstable** (σ_+) eigenvalues of A:

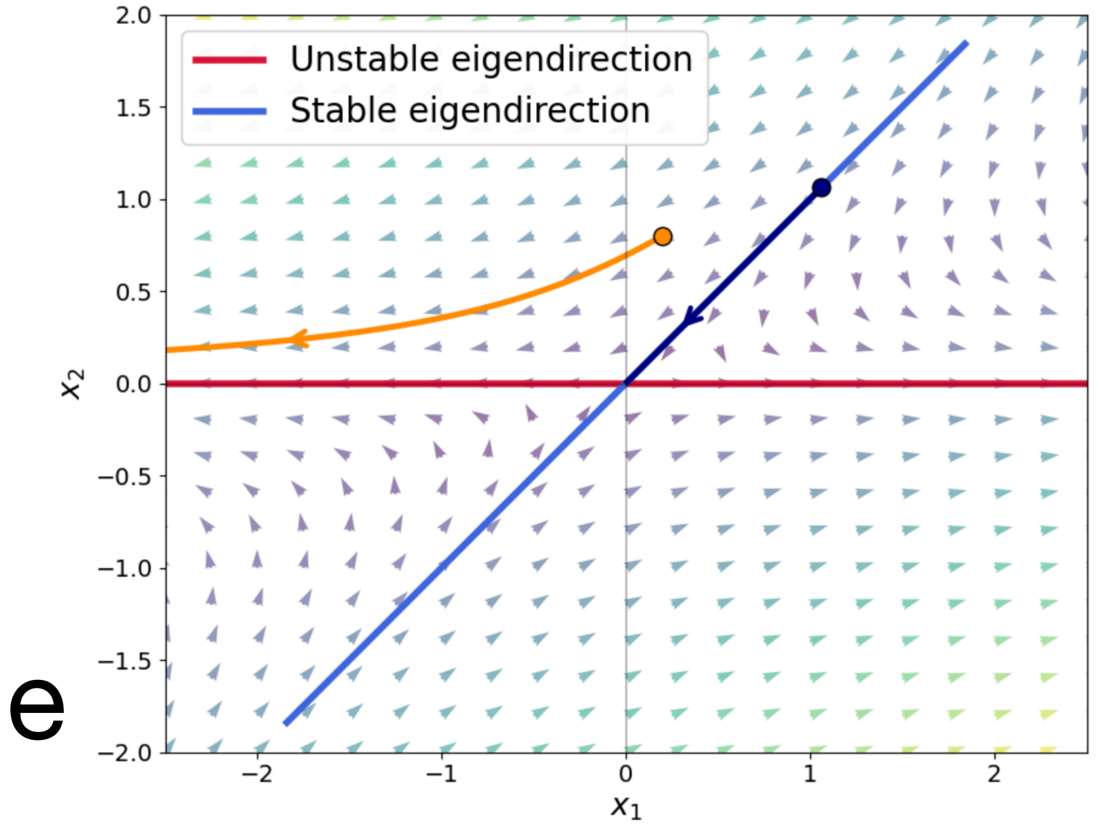
$$\sigma_-(A) := \{\lambda \in \sigma(A) \mid |\lambda| < 1\}, \quad \sigma_+(A) := \{\lambda \in \sigma(A) \mid |\lambda| \geqslant 1\}.$$

Idealized Encoder for Linear Systems

- Let us consider an ideal coordinate transform $(z, \eta) = (Ex, Nx) := Sx$, with:

$$\mathrm{range}(E^\top) = \mathcal{G}_+(A^\top),$$ [Generalized (real) eigenspaces of $A^\top$]

$$\mathrm{range}(N^\top) = \mathcal{G}_-(A^\top).$$ [Assume latent dimension $\ell = \dim(\mathcal{G}_+(A^\top))$]



- Since both $\mathcal{G}_\pm(A^\top)$ are $A^\top$-invariant and $\mathbb{R}^d = \mathcal{G}_+(A^\top) \oplus \mathcal{G}_-(A^\top)$, the (z, η)-dynamics are **block-diagonal**:

- $$\begin{bmatrix} z_{t+1} \\ \eta_{t+1} \end{bmatrix} = \begin{bmatrix} A_z & 0 \\ 0 & A_\eta \end{bmatrix} \begin{bmatrix} z_t \\ \eta_t \end{bmatrix} + \begin{bmatrix} B_z \\ B_\eta \end{bmatrix} u_t, \quad A_z := EAE^\dagger,\ A_\eta := NAN^\dagger,$$

$$B_z := EB,\ B_\eta := NB.$$

- Also, the spectrum is **preserved**: $\sigma(A_z) = \sigma_+(A)$, $\sigma(A_\eta) = \sigma_-(A)$.

Idealized Encoder for Linear Systems

- **Claim:** (A, B) is stabilizable $\Leftrightarrow$ (A_z, B_z) is stabilizable.

- $\Rightarrow$ **direction:** PBH test for stability.

- $\Leftarrow$ **direction:** Let K_z stabilize (A_z, B_z), i.e., $A_z + B_z K_z$ is Schur stable.

  - Plugging $u_t = K_z z_t$ back into (A, B):

$$\begin{bmatrix} z_{t+1} \\ \eta_{t+1} \end{bmatrix} = \begin{bmatrix} A_z + B_z K_z & 0 \\ B_\eta K_z & A_\eta \end{bmatrix} \begin{bmatrix} z_t \\ \eta_t \end{bmatrix}$$

[Zero dynamics]

  - Spectrum of the closed-loop matrix is **stable**: $\sigma(A_z + B_z K_z) \cup \sigma(A_\eta)$

Learning Stable Zero Dynamics

- With target $\mathrm{range}(E^\top) = \mathcal{G}_+(A^\top)$, let's turn back to the **multi-step loss:**

$$J_{\mathrm{bwd}}(E, D, \Phi_z) := \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_{(x,\mathbf{u})}\|\Phi^t(x, \mathbf{u}) - D(\Phi_z^t(E(x), \mathbf{u}))\|^2.$$

- **Simplifications for analysis:**
    - Autonomous flows: $\mathbf{u} = 0$.
    - Linear dynamics: $\Phi^t(x,0) = A^t z$, $\Phi_z^t(z,0) = L^t z$.
    - Linear encoder/decoder: $E(x) = Ex$, $D(x) = Dx$.
    - Isotropic initial condition: $\mathbb{E}[x] = 0$, $\mathbb{E}[xx^\top] = I$.

Learning Stable Zero Dynamics

- Simplified linear dynamics backwards loss:

$$J_{\mathrm{bwd}}(E, D, L) := \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_{x\sim \mathsf{N}(0,I)} \|A^t x - DL^t E x\|^2 = \frac{1}{T}\sum_{t=0}^{T-1} \|A^t - DL^t E\|_F^2.$$

- **Low-rank matrix factorization**, with consistency across matrix powers.
  - *Orbit* of solutions: $J_{\mathrm{bwd}}(E, D, L) = J_{\mathrm{bwd}}(M^{-1}E, DM, M^{-1}LM)$ for $M \in \mathrm{GL}(\ell)$.

- **Goal:** Characterize optimal solutions $(E_\star, D_\star, L_\star) \in \mathrm{argmin}_{E,D,L} J_{\mathrm{bwd}}(E, D, L)$.

Learning Stable Zero Dynamics

- **Theorem** [LLT26, to appear]: Any optimal solution $(E_\star, D_\star, L_\star)$ satisfies:
  - $\alpha_T \| P_{\mathcal{G}_+(A)} - P_{D_\star} \|_F^2 + \beta_T \| P_{\mathcal{G}_+(A^\top)} - P_{E_\star^\top} \|_F^2 \leqslant \dfrac{C(A,d)}{T},$

    [Optimal encoder span] **[Multi-step length]**
  - where the constants (α_T, β_T) are:

$$F_T := T^{-1} \textstyle\sum_{t=0}^{T-1} A^t (A^t)^\top \succcurlyeq \alpha_T P_{\mathcal{G}_+(A)},$$

[Loewner order in unstable subspace]

$$G_T := T^{-1} \textstyle\sum_{t=0}^{T-1} (A^t)^\top A^t \succcurlyeq \beta_T P_{\mathcal{G}_+(A^\top)}.$$

- **Note:** $\alpha_T, \beta_T = \Omega(1)$ if A has marginally stable eigenvalues, $\Omega(r^T)$ if $r := \min\{\,|\lambda| \mid \lambda \in \sigma_+(A)\} > 1$.

Learning Stable Zero Dynamics

- Recall the main objective: $J_{\text{bwd}}(E, D, L) = T^{-1} \sum_{t=0}^{T-1} \|A^t - DL^tE\|_F^2$.
    - For "large" T, the unstable parts of A^t will dominate *unless* they are cancelled out.
- Recall $(z, \eta) = (Ex, Nx) =: Sx$ which block-diagonalizes A, i.e., $SAS^{-1} = \text{blkdiag}(A_z, A_\eta)$.
    - Consequently, $A^t = S^{-1}\text{blkdiag}(A_z^t, A_\eta^t)S = (S^{-1})_1 A_z^t E + (S^{-1})_2 A_\eta^t N$.
- Setting $DL^tE = (S^{-1})_1 A_z^t E$, the **un-modeled** error is:
    - $J_{\text{bwd}}(E, (S^{-1})_1, A_z) = T^{-1} \sum_{t=0}^{T-1} \|(S^{-1})_2 A_\eta^t N\|_F^2 = O(T^{-1})$ since $\sigma(A_\eta) = \sigma\_(A)$.

- Hence $(E, (S^{-1})_1, A_z)$ is a $O(1/T)$-**approximate solution**— this yields a bound on the optimal $(E_\star, D_\star, L_\star)$ via the restricted eigenvalue (α_T, β_T) constants.

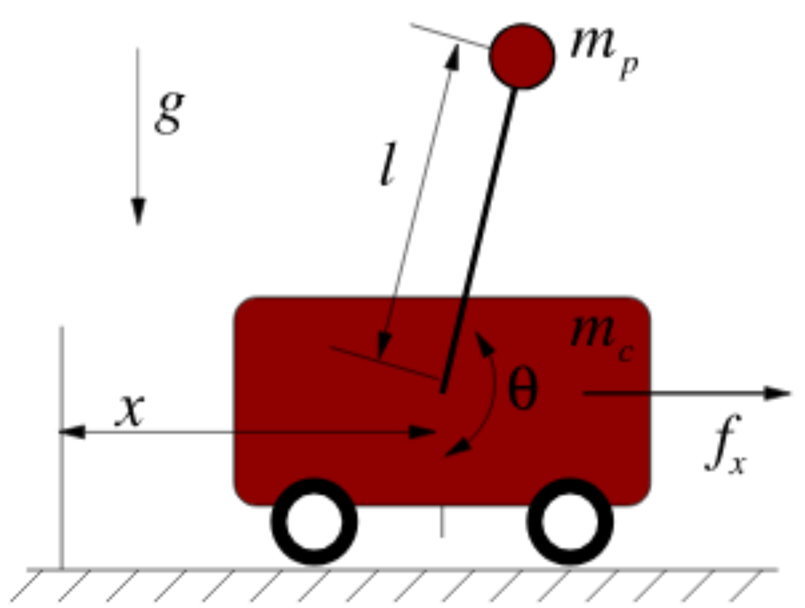
Learning Stable Zero Dynamics

- **Latent dynamics recovery:** Suppose that $D_0^\mathsf{T} D_\star$ is invertible, where $\mathrm{range}(D_0) = \mathscr{G}_+(A)$. Then,

$$\gamma_{T-1} \inf_{M \in \mathrm{GL}(\ell)} \|A_z - M L_\star M^{-1}\|_F^2 \leqslant \frac{C(A, d, E_\star, D_\star, L_\star)}{T},$$

- where $\gamma_{T-1} := \lambda_{\min}\left((T-1)^{-1} \sum_{t=0}^{T-2} A_z^t (A_z^t)^\mathsf{T} \right)$.

Back to the Cartpole Example



https://underactuated.csail.mit.edu/

- Jacobian linearization about the **upright equilibrium**:

$$\begin{bmatrix} \dot{x} \\ \dot{\theta} \\ \ddot{x} \\ \ddot{\theta} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ \theta \\ \dot{x} \\ \dot{\theta} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} u.$$

- Eigenvalues of $I + \tau A$:

$$1,\ 1,\ 1+\sqrt{2}\tau,\ {\color{blue}1-\sqrt{2}\tau} \implies \dim(\mathcal{G}_+(A^\top)) = 3,\ \dim(\mathcal{G}\_(A^\top)) = 1.$$

- Hence why encoding $E(x) : \mathbb{R}^4 \mapsto \mathbb{R}^3$ worked for stabilization.

Conclusions

Summary

- We built some insight into how certifying stability on closed-loop latent dynamics can yield stable closed-loop behavior on the original system.

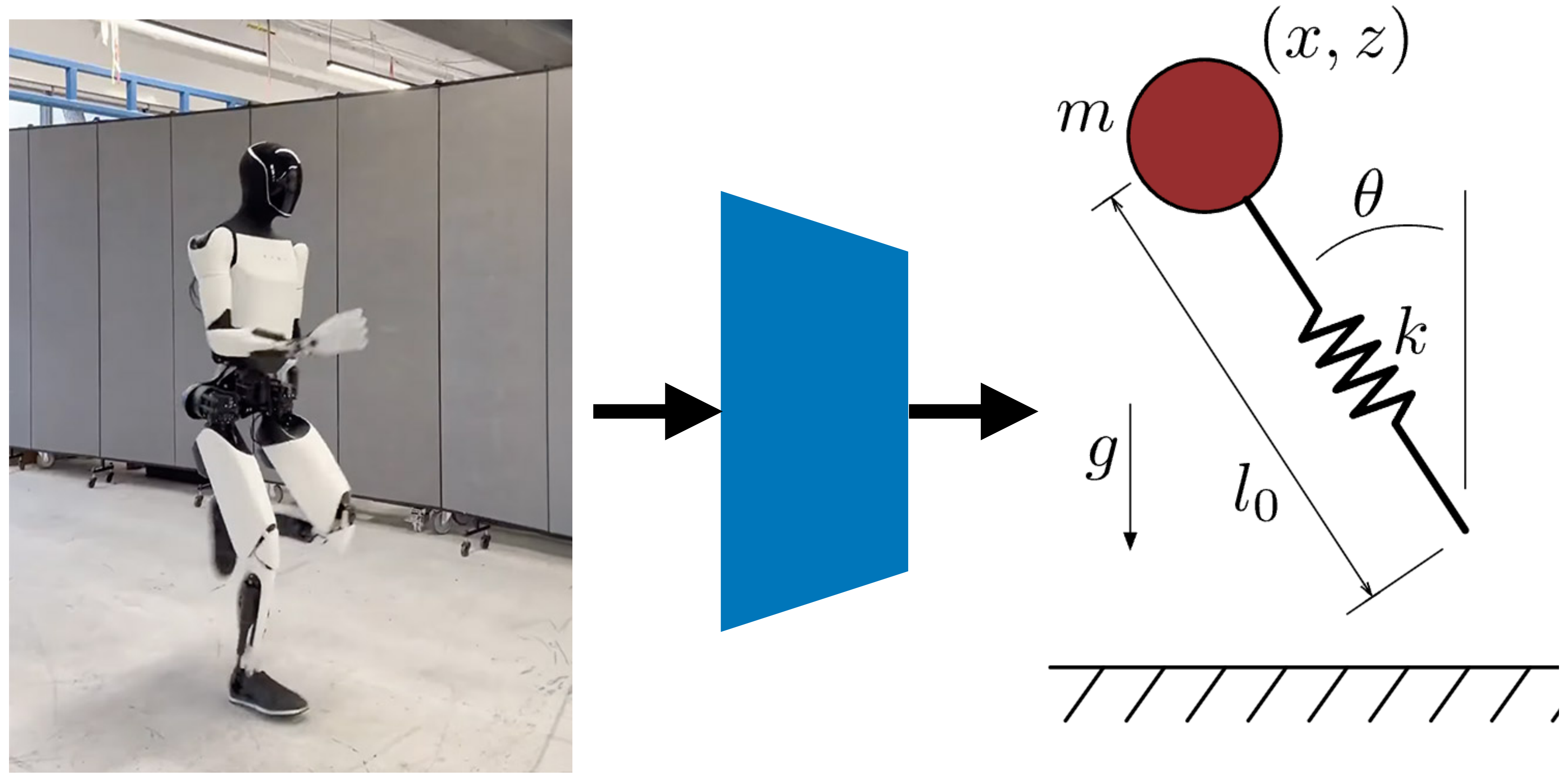
- A lot left to do to make control from latent variable models a rigorous tool:

    - Learning zero-dynamics for **non-linear** systems:

        - Local convergence (via Jacobian linearization or spectral sub-manifolds)

        - Global convergence (via Koopman operators).

    - Incorporating **vision/sensory** feedback.



Observations

Latent Model

Thanks!

stephentu.github.io
stephen.tu@usc.edu

Why $\mathcal{G}_{\pm}(A^\top)$ Instead of $\mathcal{G}_{\pm}(A)$?

- Our goal is to write $z_{t+1} = A_z z_t$ and $\eta_{t+1} = A_\eta \eta_t$ (forget the inputs for now).
- This means we want $z_{t+1} = Ex_{t+1} = EAx_t$ equal to $A_z E x_t = A_z z_t$. Thus:

$$
\begin{aligned}
EA = A_z E \text{ for some } A_z &\iff A^\top E^\top = E^\top A_z^\top \text{ for some } A_z \\
&\iff \mathrm{range}(E^\top) \text{ is } A^\top\text{-invariant} \\
&\Longleftarrow \mathrm{range}(E^\top) = \mathcal{G}_{+}(A^\top).
\end{aligned}
$$