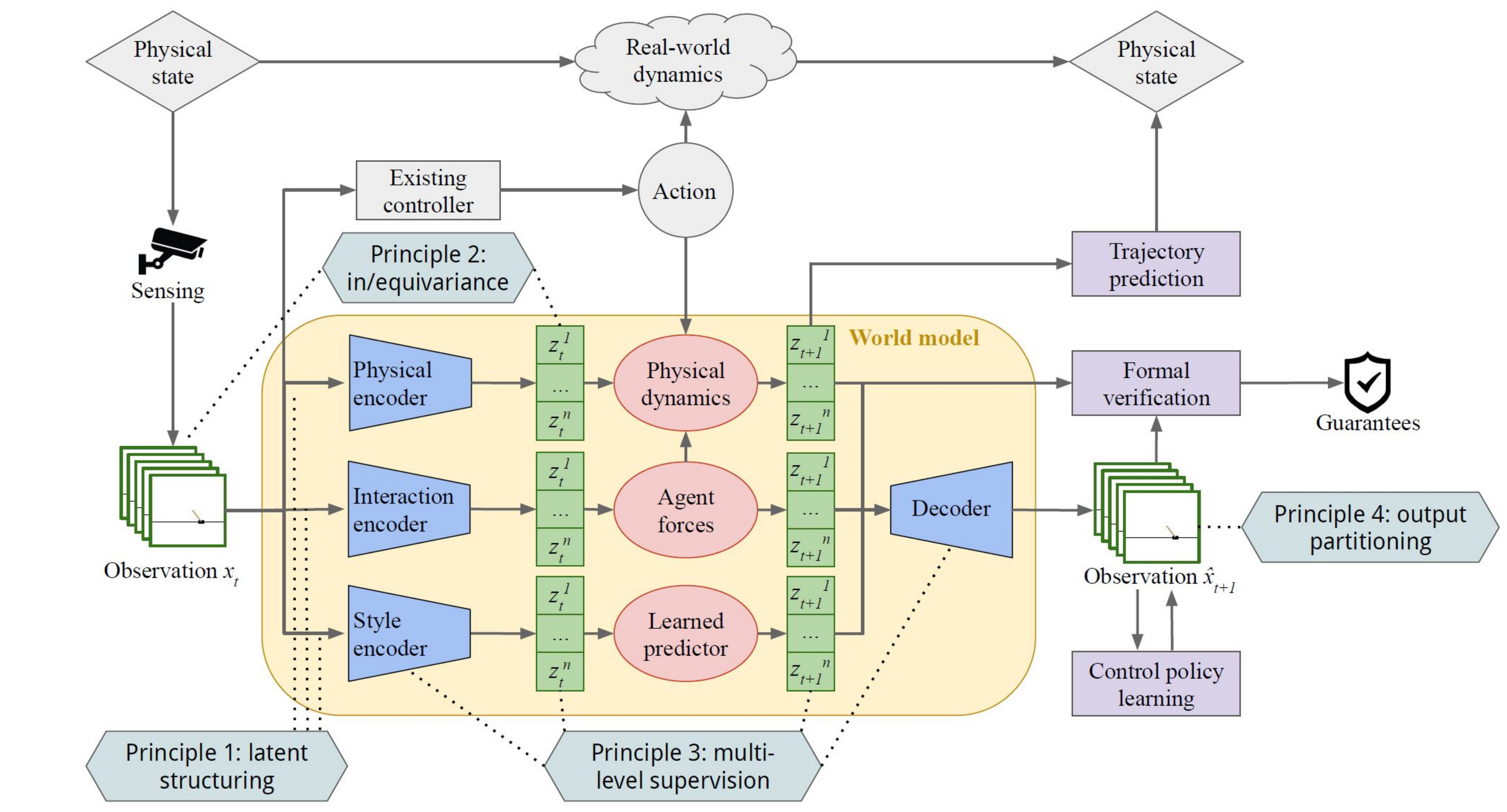
# Four Principles for Physically Interpretable World Models


Jordan Peper*, Zhenjiang Mao*,
Yuang Geng, Siyuan Pan, Ivan Ruchkin

*Trustworthy Engineered Autonomy (TEA) Lab*

*University of Florida*


## PROBLEM & CONTRIBUTIONS




**What is a world model?** Let $f(x) = (\mathrm{dec} \circ \mathrm{dyn} \circ \mathrm{enc})(x)$ be a *world model*:

- enc: $X \rightarrow Z$ encodes the observation $x$ into latent space
- dyn: $Z \rightarrow Z$ propagates the latent embedding through time
- dec: $Z \rightarrow X$ decodes the embedding back into the observation space

**Problem:** latent space lacks **physical interpretability**, making it difficult to:

- Understand what the model "knows"
- Integrate classical, state-based controllers or planners
- Provide physically grounded safety guarantees

**Solution:** train world models that are **physically interpretable**

- Latent embeddings $z$ correspond to physical properties
- Latent dynamics dyn emulate physical processes

TEA Lab:

Paper:

## PRINCIPLE 1

### Functionally organized latent space

- **Principle:** functionally organize the latent space
  - — Modular latent embedding and dynamics to encode human conceptual priors (absolute agent dynamics, relative dynamics between other agents, and background features)
- Overall loss is proportional to the losses in each branch:

$$\mathcal{L} \propto L_1(f_1(\mathrm{enc}_1(x)), x) + L_2(f_2(\mathrm{enc}_2(x)), x) + L_3(f_3(\mathrm{enc}_3(x)), x)$$

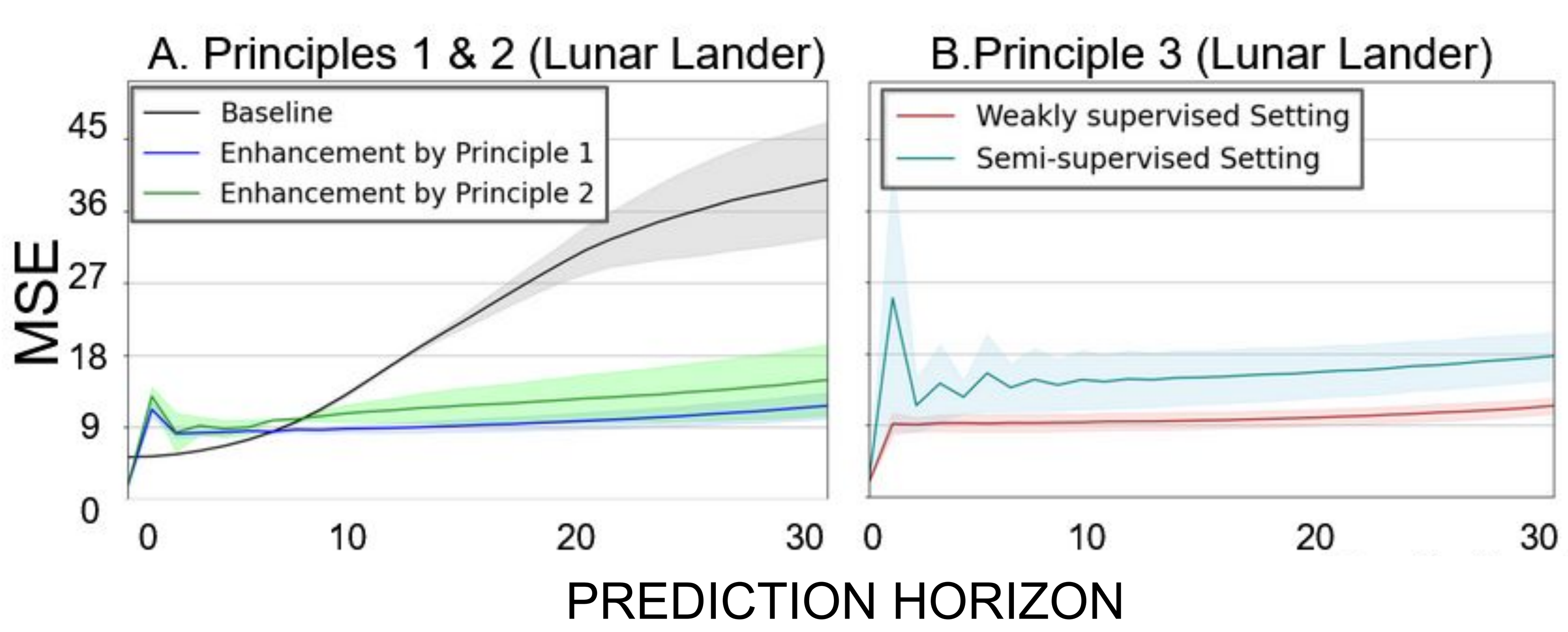
- *L* = loss fn, $f_i$ = WM branch, *enc* = encoder, *x* = input/observation

## PRINCIPLE 3

### Multi-Level and Multi-Strength Supervision

- **Principle**: integrate supervision signals of varied strength from multiple abstraction levels
- *Key Idea:* Physical supervision signals vary in both form (e.g., states, trajectories, constraints) and strength (e.g., exact values, intervals, implicit patterns). Training should adapt accordingly.
- *Why It Matters:* Real-world data often includes a mix of precise labels, coarse annotations, and entirely unlabeled sequences.

## EXPERIMENTAL RESULTS




- Functionally organizing the latent space by physical roles (e.g., dynamics, interaction, style) **improves stability** over long horizons
- Encoding physical symmetries into the latent space **enhances generalization** to transformed observations
- Given partially labeled data, adding extra physical signals (e.g., inferred velocity or constraints) significantly **improves** learning and long-term **prediction**


**Reference**: *Peper, J.*, Mao, Z.*, Geng, Y., Pan, S., & Ruchkin, I. (2025). Four Principles for Physically Interpretable World Models. In Proceedings of the International Conference on Neuro-symbolic Systems (NeuS), Philadelphia, PA, 2025. *Equal contribution*


## PRINCIPLE 2

### Invariant/equivariant representations

- **Principle:** learn invariant/equivariant representations of the environment
  - — *Invariant:* **do not transform** for transformation *f* that does not affect the underlying meaning (noise, image rotation)
  - — *Equivariant:* **do transform** for transformation *f* that affects the underlying meaning (fog, shape distortion)

$$\mathcal{L}_{wm}(x) \propto \frac{\lambda}{|T|} \sum_{(g,h)\in T} \| \mathrm{enc}(g(x)) - h(\mathrm{enc}(x)) \|_2^2$$

- *x* = input/observation, *enc* = latent encoder, *g* = input-space transformation, *h* = latent-space transformation
  - — To promote invariance: *h(enc(x)) = enc(x)*

## PRINCIPLE 4

### Partitioned World Model Generation

- **Principle:** partition generated observations into segments from multiple simpler generators (enables **scalable verification**)
- Loss function: $\mathcal{L}_{gen} = ||x - \hat{x}||^2 + \lambda \sum_{i=1}^{N} ||x_i - \hat{x}_i||^2$

— **Segment-wise reconstruction loss:** Enforces each decoder to accurately model its part of the input

— **Combined reconstruction loss:** Encourages the model to reproduce the full observation when segments are combined

| World model | Environment | Average MSE | Average SSIM | Model Size |
|---|---|---|---|---|
| Baseline (monolithic) | Cart Pole | 0.02856 | 0.997122 | 200,259 |
| Partitioned 3-way | Cart Pole | 0.05176 | 0.995614 | 144,665 |
| Baseline (monolithic) | Lunar Lander | 0.18801 | 0.8686 | 360,773 |
| Partitioned 3-way | Lunar Lander | 0.306 | 0.6289 | 78,101 |

## GAP IN EXISTING WORK