Behavioral Monitoring of JEPA World Models with Jacobian Centroids

Thomas Walker [1] Randall Balestriero [2] Richard Baraniuk [1]

[1]Rice University [2]Brown University

Motivation

Deploying World Models (WMs) for planning requires monitoring internal representations for...

- ...goal alignment.
- ...failure.
- ...out-of-distribution detection.

Activation-based signals, reflect what the model *knows*, not what it will *do*.

Centroids, obtained from sub-component Jacobians, give a *behavioral* signal that directly predicts planning failure and enables adaptive replanning in WMs.

Centroids as a Behavioral Signal

For a sub-component g of a WM and an input $\boldsymbol{x}$, the centroid is the row-sum of the input–output Jacobian:

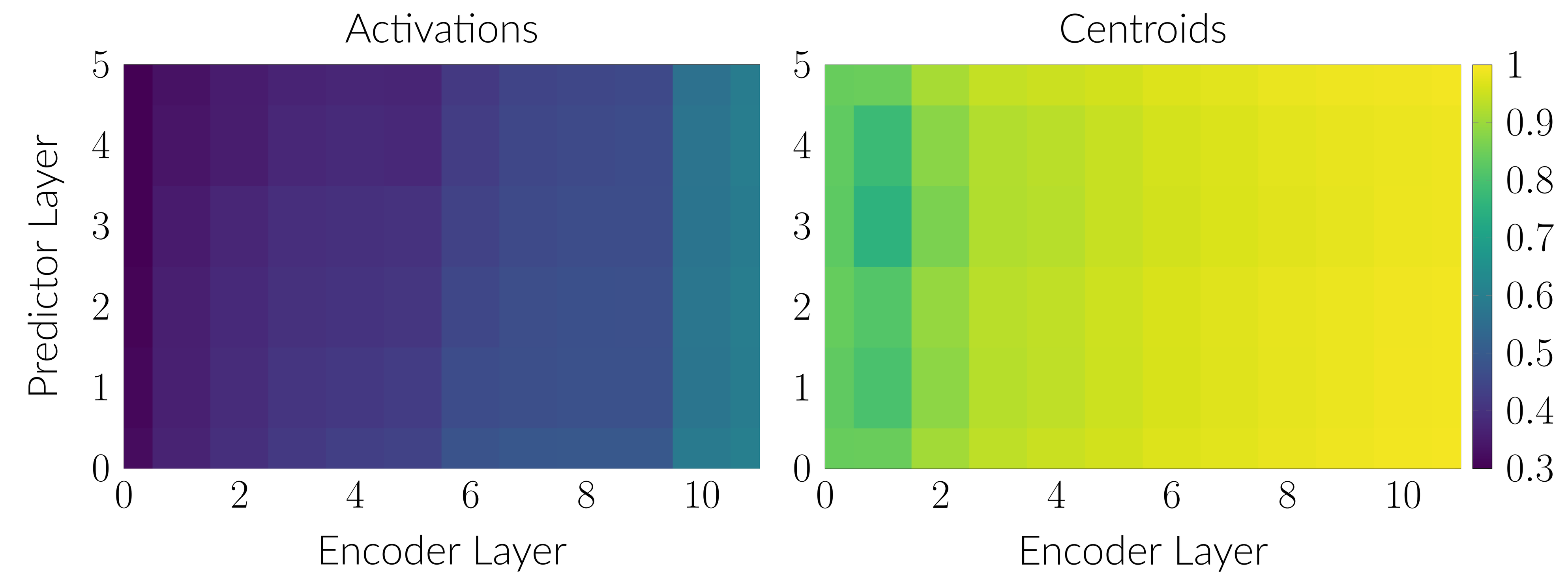
$$\boldsymbol{c} = \left(\boldsymbol{J}_g[\boldsymbol{x}]\right)^\top \mathbf{1}, \qquad \text{vs. the activation } g(\boldsymbol{x}).$$

Centroids are motivated under the **Linear Centroids Hypothesis (LCH)**, which shows that centroids have a linear structure that captures model behavior.

We study the centroids and activations of encoder-predictor **LeWorldModels**.

The Predictor Tracks the Encoder Behaviorally

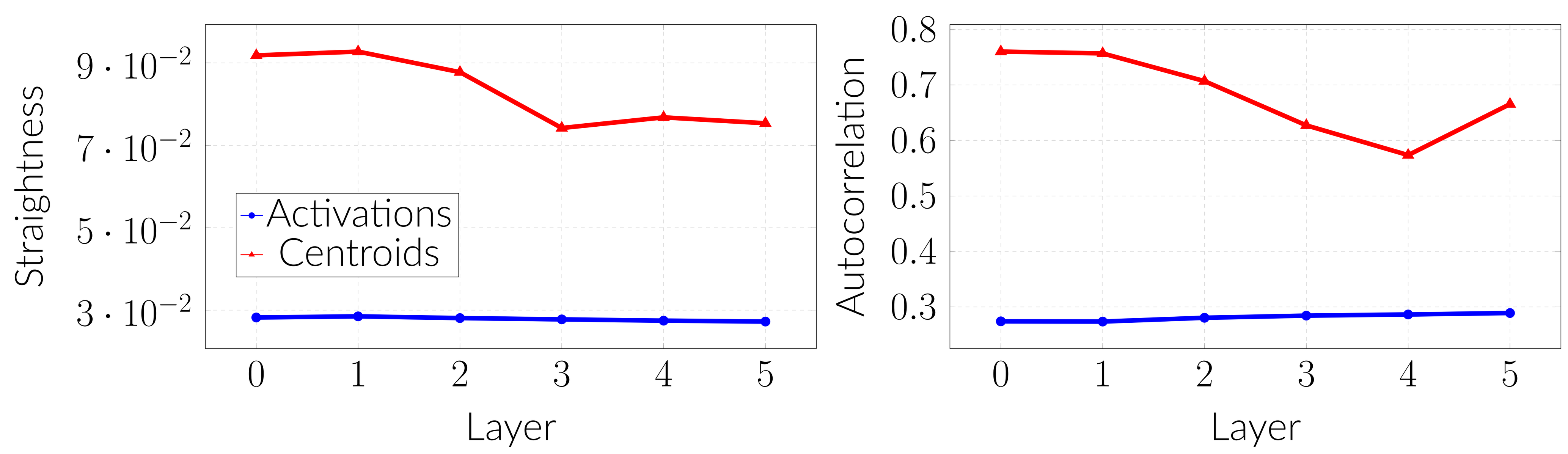
Ridge regression between layer-wise PCA projections of centroids/activations shows each predictor layer's behavioral sensitivity is linearly predictable from the encoder's — more strongly so for centroids.



Expert Trajectories Are Smooth in Behavioral Space

Compared to activations, centroids are consistently...

- ...*straighter* (end-to-end displacement / path length).
- ...*correlated* (lag-1 autocorrelation).



Centroids Predict Failed Trajectories

Can an internal signal tell us whether an episode is heading toward failure?

We compare centroid signals against activation signals, on Push-T and TwoRoom. In-distribution sets the goal offset to 25 steps; out-of-distribution sets it to 75 steps.

- **Calibration**: cosine similarity between realized and predicted next-step centroid/embedding (c_{cen} vs. c_{act}).
- **SAE Goal-code**: Using a TopK-SAE dictionary of atoms built from 2000 expert trajectories (sae-cen vs. sae-act).

Signal	In-dist.		Out-of-dist.	
	Push-T	TwoRoom	Push-T	TwoRoom
c_{act} (calibration)	**0.88**	0.76	0.91	0.82
c_{cen} (calibration)	0.86	0.91	0.91	0.91
c_{goal} sae-act	0.71	0.91	0.95	0.97
c_{goal} sae-cen	0.68	**0.92**	**0.95**	**0.98**

Encoder–Predictor Dissociation Explains Failure

Where do planning trajectories break down?

We split trajectories on encoder-centroid and predictor-centroid alignment. We identify associated and dissociated trajectories.

1. **Associated trajectories**: the encoder *correctly represents the goal* and the predictor is *behaviorally responsive*.
2. **Disassociated trajectories**: Either the encoder *misrespresents the goal* or the predictor is *behaviorally unresponsive to it*.

Centroid axes isolate this dissociation far more cleanly than activation axes (r=0.58 vs. r=0.19).

	Push-T	TwoRoom
Associated	45.6%	89.7%
Dissociated	6.4%	9.9%

Associated episodes succeed far more often out-of-distribution.

A Gate for Adaptive Replanning

Can a doomed plan be flagged and recovered?

Among dissociated episodes, only those in which the encoder remains aligned are *recoverable*.

- **Gate**: threshold predictor-centroid cosine at the first CEM planning call, calibrated *without labeled failures*.
- **Flag**: before next actions are taken, mark episodes below the threshold.
- **Redirect**: flagged episodes are resampled to a closer goal.

Task	n flagged	Original SR	Resampled SR
Push-T	72 (29%)	0.0%	26.4%
TwoRoom	125 (50%)	3.8%	29.6%