Video foundation models learn motion, not physics, through structured temporal representations.

YORK U

VECTOR INSTITUTE | INSTITUT VECTEUR

Mila

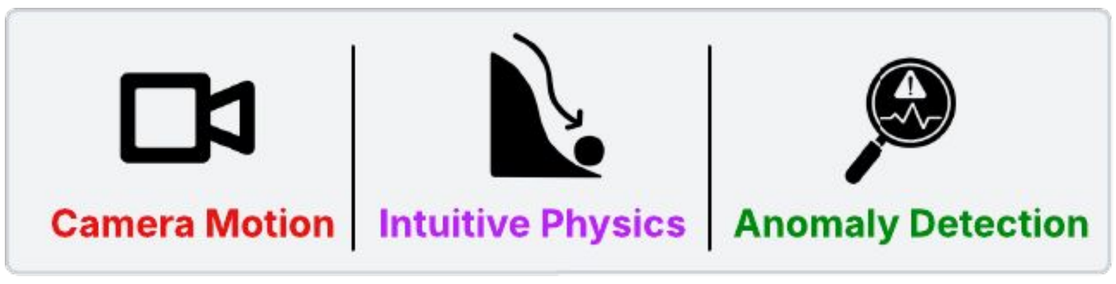
Meta

What, Where, and How: Probing Spatiotemporal Representations in Video Foundation Models

Sharon S. Musa, Fereshteh Forghani, Harrish Thasarathan, Sonia Joseph, Matthew Kowal, Konstantinos G. Derpanis

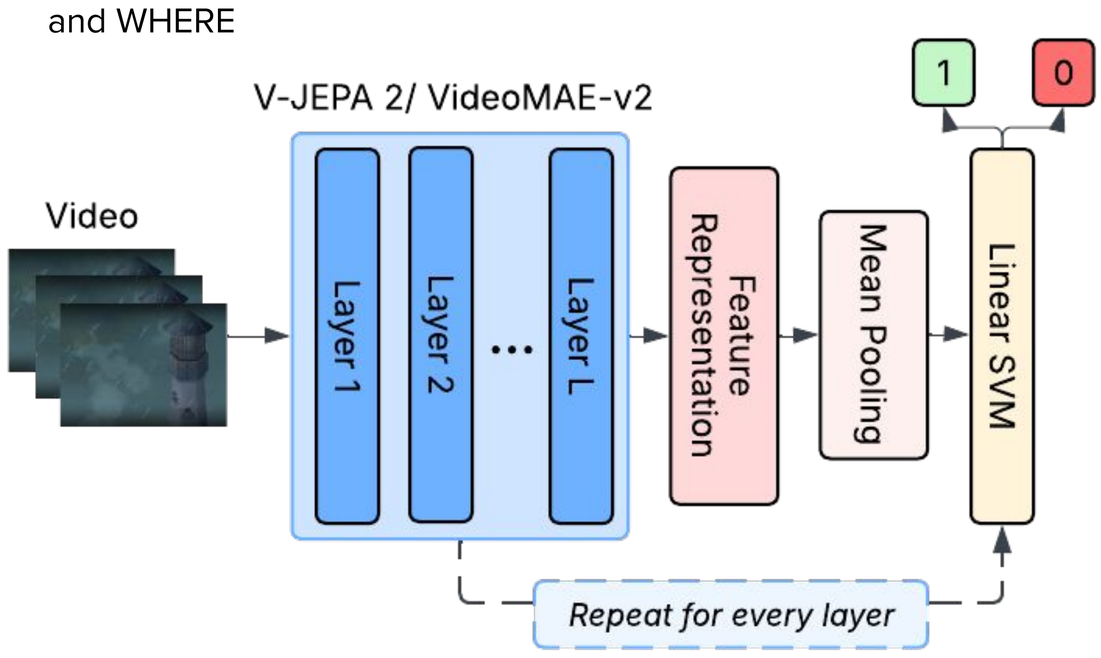
1 BACKGROUND AND MOTIVATION

- Videos contain rich temporal signals that static images cannot capture.
- We study what temporal properties are encoded in video foundation models, where they emerge across layers, and how they are geometrically organized.
- Properties evaluated include Camera Motion. Intuitive Physics and Anomaly Detection.

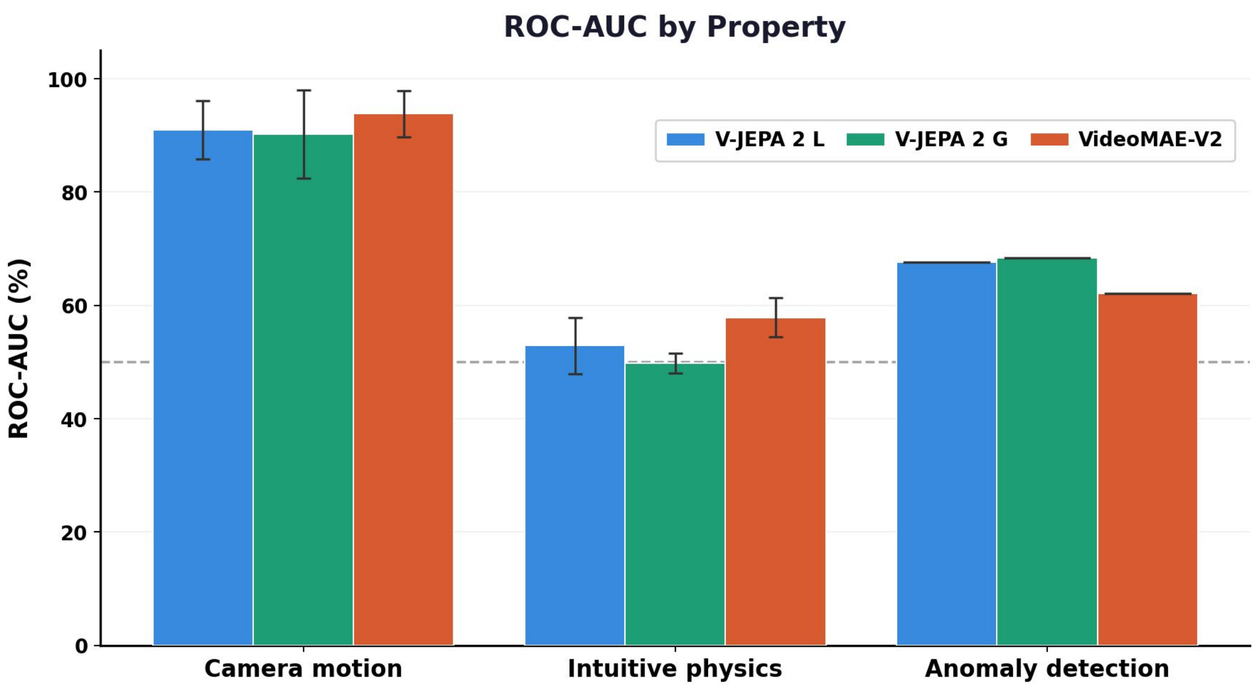


2 Layerwise Probing

- We extract features from every transformer layer of V-JEPA 2 and VideoMAE-v2.
- Train seperate lightweight probes at each layer
- Evaluate on the three task families/ properties to answer WHAT and WHERE



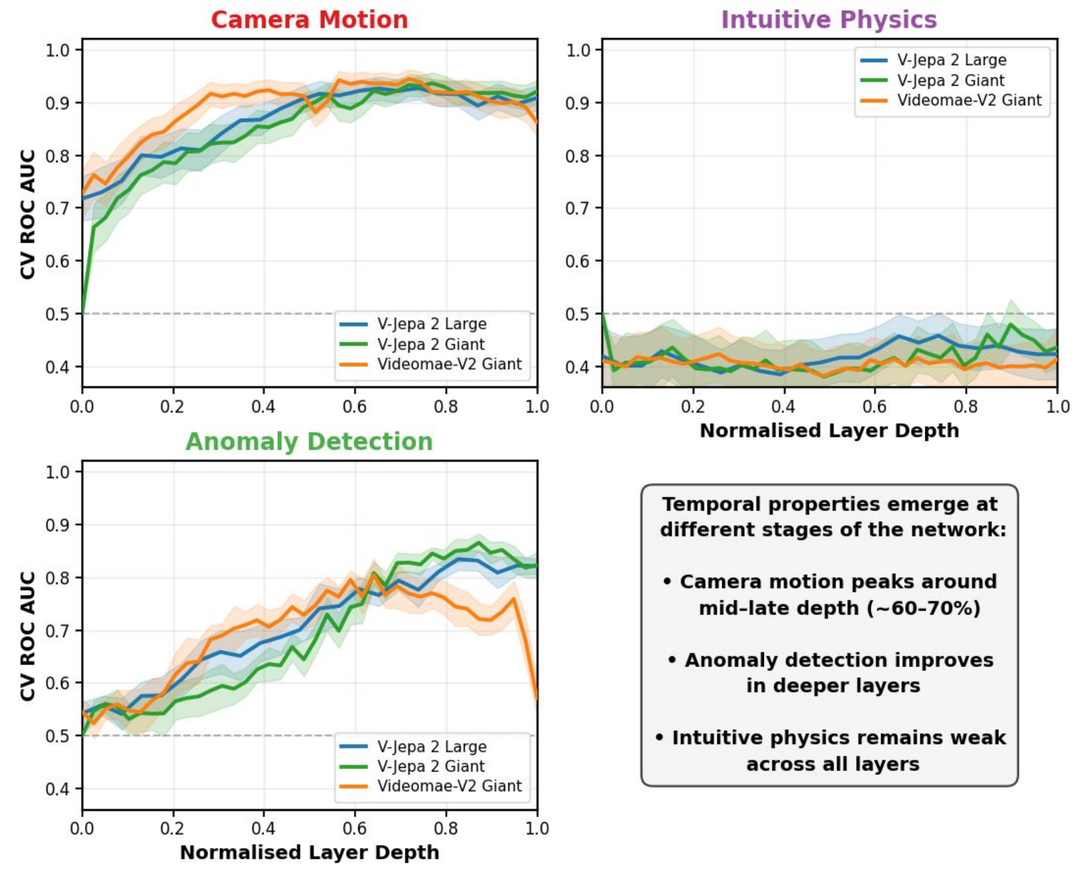
3 WHAT: what is encoded



Models strongly encode motion cues, but show little evidence of intuitive physics understanding

- Camera motion: >90 ROC-AUC across all models
- Anomaly detection: moderate performance (>60 ROC AuC)
- Intuitive physics: near chance

4 WHERE: where is it encoded?

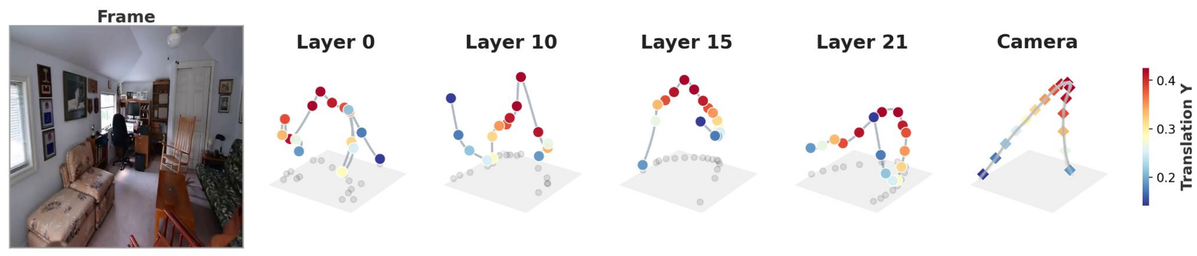


Temporal properties emerge at different stages of the network:

- **Camera motion peaks around mid–late depth (~60–70%)**
- **Anomaly detection improves in deeper layers**
- **Intuitive physics remains weak across all layers**

5 HOW: how is it structured?

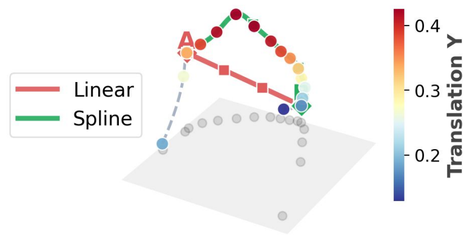
- Temporal representations lie on smooth manifolds in latent space, enabling geometry-aware interventions.
- Smooth 1d trajectories emerge in mid-to-late layers
- Camera motion changes continuously along these curves



6 Steering

- We Interpolate between two frames, retrieving the nearest real neighbour
- Linear steering cuts through space and jumps across frames
- Spline steering follows the learned trajectory resulting in smoother temporal progression

Therefore, following the learned geometry yields more coherent videos



7 Limitations and Discussion

- Our results show what is linearly decodable, Nonlinear probes may reveal additional structure, especially for intuitive physics, though at the cost of reduced interpretability.
- IntPhys2 and UBnormal use synthetic videos, whose visual statistics may differ from the natural videos used for pretraining.

Demo page

smusa28@my.yorku.ca