V-JEPA 2 encodes motion direction causally from very early layers, exhibiting functional subspace stability even when its geometrics embedding shifts.

Causal State Variables in V-JEPA 2 Latents: Discovery, Intervention, and Portability

Background: JEPAs achieve strong motion-understanding performance, but linear probing cannot tell us if their latents are causally functional. We applied a 3-stage causal intervention to a frozen V-JEPA 2 encoder to find out.

Result 1: Early & Distributed Causal Encoding

V-JEPA 2 encodes motion direction causally from early layers, distributed across a mjority of latent dimensions. Causal ablation at layer 7 is 43× more specific than random-direction controls.

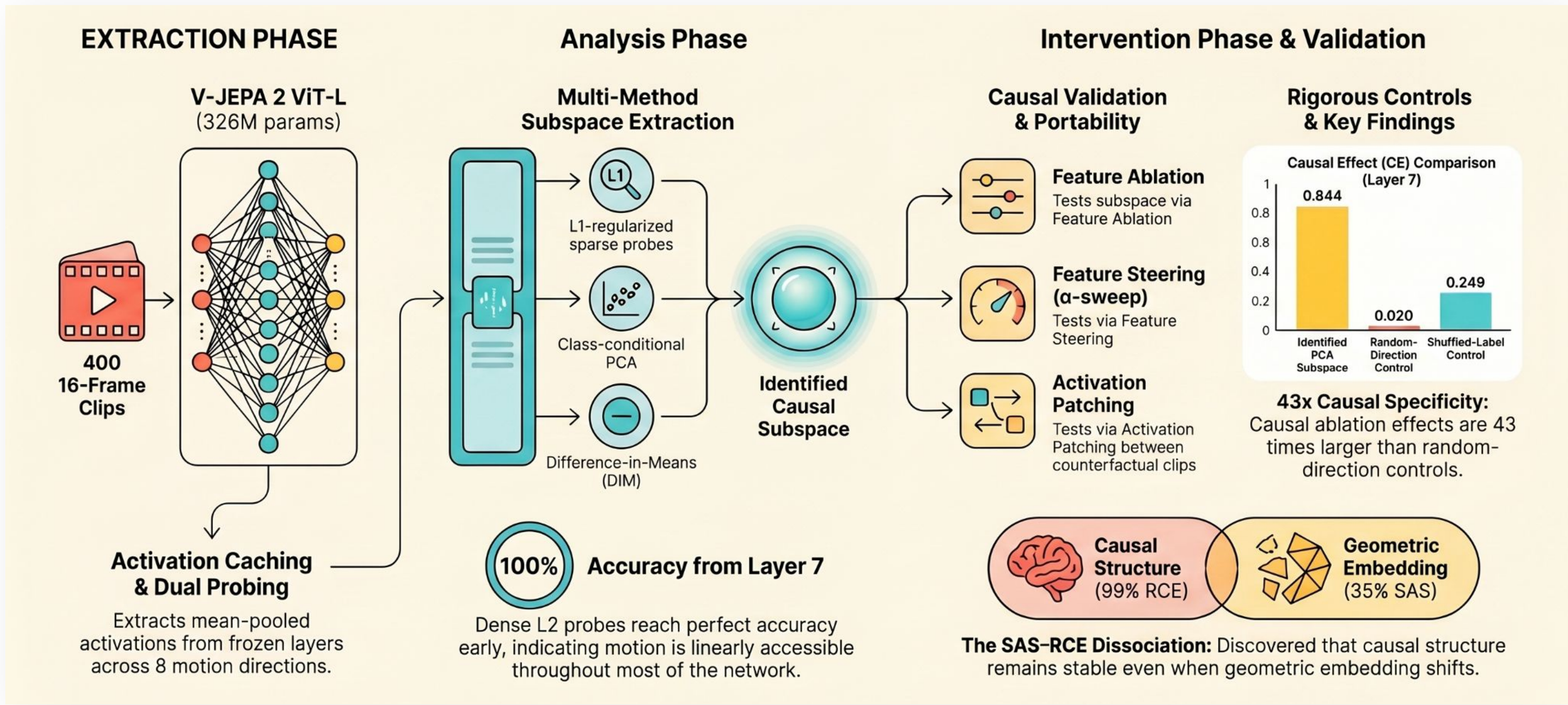
Metric	Proxy (d=256)	V-JEPA 2 (d=1024)
Best dense accuracy	100% (L18–23)	100% (L7–20)
Onset layer	L14 (88.8%)	L4 (96%)
Sparse dims (ρ)	84/256 (ρ=0.33)	585/1024 (ρ=0.57)
CE (ablation)	0.814	0.844
CE (random control)	0.001	0.020
CE ratio	818×	43×
PSR	0.99	0.88
Early-layer CE	0.025	0.815
SAS (PCA)	0.65	0.35
RCE (PCA)	0.35	0.99

Result 2: The Geometric-Functional Dissociation

Causal structure is far more stable than its geometry. Across input perturbations, Subspace Alignment is moderate, but Retained Causal Effect remains near-perfect. This holds across all tested features.

Experiment	Model	Best Acc.	CE Ratio	PSR	SAS / RCE
SCS (simple, 8-way)	ViT-L	100% (L7)	43×	0.88	0.35 / 0.99
CSC (complex, 8-way)	ViT-L	100% (L8)	2.7×†	0.67	0.32 / 0.96
Kinetics (real, 4-way)	ViT-L	100% (L4)	5.3×	0.37‡	0.46 / 1.30
SCS (PCA-100d)	ViT-L	100% (L6)	13×	0.78	—
SCS (PCA-100d)	ViT-H	100% (L9)	54×	0.70	—
CCA cross-model (L→H)		—	—	—	1.00 / 1.00

Methods



Limitation: Interventions establish causal sufficiency of the extracted subspaces for a downstream readout, but do not provide a full circuit-level mechanism linking earlier individual attention heads or MLPs to this formation.

PAPER

Guus Bouwens

Mechanistic Interpretability Workshop, ICML, Seoul, South Korea, 2026

ICML
International Conference
On Machine Learning