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

Anonymous Authors[1]

Abstract

Video world models trained with Joint-Embedding Predictive Architectures (JEPAs) achieve strong performance on motion-understanding benchmarks, but whether their latent representations encode *causally functional* state variables remains unknown. We apply a three-stage causal-state discovery pipeline—combining L1-regularized probing, class-conditional PCA, difference-in-means subspace extraction, and three families of causal interventions with four matched controls—to the frozen encoder of V-JEPA 2 ViT-L (326M parameters, d=1024, 24 layers) on a synthetic controlled-sequence dataset of 400 clips across 8 motion directions. V-JEPA 2 encodes motion direction from remarkably early layers (96% dense-probe accuracy at layer 4; 100% by layer 7), using a distributed subspace occupying 57% of latent dimensions. Causal ablation at layer 7 produces effects $43\times$ larger than random-direction controls, confirming the identified subspace is causally privileged. The SAS–RCE dissociation—moderate subspace alignment (SAS $=$ 0.35) coexisting with near-perfect retained causal effect (RCE $=$ 0.99)—reveals that causal structure is far more stable than its geometric embedding. Findings generalize to complex synthetic stimuli, real Kinetics video ($5.3\times$ CE ratio), and V-JEPA 2 ViT-H ($54\times$ CE ratio with near-perfect cross-architecture CCA alignment). These results provide the first intervention-based evidence that JEPA video models encode motion as a causally functional latent variable, and introduce SAS and RCE as portability metrics for mechanistic interpretability.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

A central aspiration of self-supervised video learning is to produce models that reason about the physical world rather than merely compress statistical regularities in pixels. Joint-Embedding Predictive Architectures (JEPAs) (LeCun, 2022) pursue this goal by training an encoder to predict the *latent representations* of masked regions from visible context, bypassing pixel-space reconstruction and thereby encouraging abstract, semantic encodings. V-JEPA (Bardes et al., 2024) extends this programme to video with spatio-temporal tube masking, and achieves strong performance on motion-understanding benchmarks such as Something-Something v2 (SSv2). The more recent V-JEPA 2 (Assran et al., 2025) scales to 1.2B parameters and enables robotic planning in zero-shot settings—suggesting that JEPA representations encode genuinely causal structure about how objects move and persist over time.

Yet our mechanistic understanding of *how* these representations are organized remains thin. The dominant tool for understanding representation quality is *linear probing*: training a linear classifier on frozen features and reporting accuracy as a proxy for encoding quality. Probing reveals that information is linearly accessible, but it cannot tell us whether that information is *causally involved* in the model's behavior or merely a passive correlate. This is the central gap we address.

In this paper, we go beyond probing. We develop and apply an intervention-based pipeline for discovering *causal state variables* in JEPA-family latent spaces, and deploy it on V-JEPA 2's frozen ViT-L encoder. Our central finding is that V-JEPA 2 encodes motion direction as a causally functional latent variable from very early layers—much earlier and more distributed than predicted by language-model analogies—and that this causal structure exhibits remarkable functional stability across input perturbations even when the geometric embedding shifts substantially.

Mechanistic interpretability and vision models. The mechanistic interpretability literature has developed a rich toolkit for causal analysis of neural networks. Activation patching (Vig et al., 2020; Meng et al., 2022) isolates the computational role of individual components by replacing

activations with values from counterfactual inputs. Causal abstraction (Geiger et al., 2024) formalizes the correspondence between model internals and interpretable high-level algorithms. Sparse autoencoders (Bricken et al., 2023; Cunningham et al., 2023; Templeton et al., 2024) decompose polysemantic activation vectors into monosemantic latent features. Almost all of this work targets language models; *video world models trained with predictive objectives remain entirely unexplored* from a mechanistic standpoint.

Contributions. We make four contributions:

1. We present the first intervention-based mechanistic analysis of a JEPA video model, demonstrating that V-JEPA 2 encodes motion direction as a *causally functional* latent variable from layer 4 onward, with $43\times$ causal specificity over random-direction controls.
2. We discover a striking dissociation between geometric and functional portability: subspace alignment is moderate (SAS $= 0.35$) but retained causal effect is near-perfect (RCE $= 0.99$), revealing that the causal structure is far more stable than its geometric parameterization.
3. We validate the complete pipeline on a structured proxy model with known ground truth ($818\times$ CE ratio, PSR $= 0.99$), and introduce two portability metrics—Subspace Alignment Score (SAS) and Retained Causal Effect (RCE)—for quantifying subspace transfer.
4. We extend the analysis to visually complex stimuli, real video, and V-JEPA 2 ViT-H, demonstrating that causal motion encoding and the SAS–RCE dissociation generalize across stimulus complexity and model scale.

2. Background and Related Work

JEPA architecture. The Image-based JEPA (I-JEPA; Assran et al. 2023) trains a context encoder f_θ and a target encoder $\bar{f}_\theta$ (updated by exponential moving average) alongside a predictor g_ϕ. Training minimizes

$$\mathcal{L}(\theta, \phi) = \sum_t \left\| \hat{\mathbf{z}}_t - \bar{f}_\theta(\mathbf{x}_{B_t}) \right\|_2^2, \tag{1}$$

where $\hat{\mathbf{z}}_t = g_\phi(\mathbf{z}_c, \text{pos}(B_t))$ predicts target block encodings from context. Crucially, the prediction target is in *latent space*, not pixel space. V-JEPA (Bardes et al., 2024) extends I-JEPA to video with temporal masking; V-JEPA 2 (Assran et al., 2025) further introduces action-conditioned fine-tuning, achieving state-of-the-art SSv2 accuracy of 77.3%.

Mechanistic interpretability methods. *Activation patching* (Vig et al., 2020; Meng et al., 2022) replaces activations from a source input with those from a counterfactual and measures downstream effects. Subspace patching (Makelov et al., 2023) generalizes this to low-rank subspaces; Makelov et al. show that naive subspace swaps can activate dormant parallel pathways, yielding illusory localization signals. We guard against this with explicit specificity controls. *Causal abstraction* (Geiger et al., 2024) provides a formal framework for when a high-level causal model faithfully describes a network's computation; we ground our interventions in this framework. SAEs have been applied to extract monosemantic features in LLMs and more recently in vision models (Pach et al., 2025; Stevens et al., 2025); we draw on SAE methodology for our sparse probe design.

Gap. No prior work applies intervention-based mechanistic analysis to JEPA latent spaces. The closest related efforts are probing studies on V-JEPA (Bardes et al., 2024) (reporting linear probe accuracy but not causal interventions) and SAE-based studies on ViTs (targeting supervised or contrastive models, not predictive world models). We fill this gap with the first causal-intervention analysis of a JEPA video encoder.

3. Method

3.1. Experimental Setup

Target model. We analyze V-JEPA 2 ViT-L/16 (Assran et al., 2025), a 326M-parameter video encoder with 24 transformer layers, hidden dimension d=1024, and 16 attention heads. The model processes 16-frame clips at 256×256 resolution with patch size 16 and tubelet size 2. We use the frozen pretrained encoder without any fine-tuning.

Structured proxy model. To validate pipeline sensitivity before interpreting V-JEPA 2 results, we construct a **structured proxy** mirroring V-JEPA's architecture at reduced scale (d=256) with an injected ground-truth signal: motion direction is embedded into late-layer activations (layers 12+) via a controlled signal with tunable SNR, while early layers produce random activations. This enables us to test whether the pipeline correctly recovers known causal structure and whether controls correctly reject null subspaces.

Datasets. *Synthetic controlled sequences (SCS):* 400 clips across 8 discrete motion directions ($\theta \in \{0°, 45°, \ldots, 315°\}$, 50 clips per direction), each showing a bright circle moving at constant velocity against a dark background for 16 frames at 256×256. *Complex stimulus control (CSC):* 200 clips (25 per direction) with textured backgrounds, 2–5 distractor objects, non-uniform motion, and per-frame brightness jitter—testing robustness to visual clutter. *Kinetics real video:* 95 clips from Kinetics-400 (Kay et al., 2017)

spanning diverse action categories; 4-way motion labels assigned via dense optical flow (Farneback, minimum 1.5 px/frame threshold). *Cross-checkpoint:* Both ViT-L (326M, d=1024, 24 layers) and ViT-H (600M, d=1280, 32 layers) process the same 400 SCS clips, with PCA reduction to 100 dimensions for comparability.

Figure 1 provides a pipeline overview.

3.2. Probing and Subspace Extraction

We extract mean-pooled activations $\bar{\mathbf{h}}^{(l)} = \frac{1}{N}\sum_i \mathbf{h}_i^{(l)}$ at every layer and train two probe types: (1) a **dense linear probe** (L2-regularized logistic regression), and (2) a **sparse probe** (L1/LASSO), yielding a sparse weight vector with $k \ll d$ non-zero entries.

From layers with highest probe accuracy, we extract candidate causal subspaces via three complementary methods: **(a) Sparse probe directions:** the non-zero entries of the L1 probe define $\hat{\mathbf{u}} = \mathbf{w}_{\text{sp}}^{(l)}/\|\mathbf{w}_{\text{sp}}^{(l)}\|_2$. **(b) Class-conditional PCA:** PCA on the matrix of class-mean differences $\mathbf{M} \in \mathbb{R}^{C\times d}$ (rows are $\boldsymbol{\mu}_c - \bar{\boldsymbol{\mu}}$) extracts the top-$r$ principal directions. **(c) Difference-in-means (DIM):**

$$\mathbf{u}_{\text{DIM}} = \frac{\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0}{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\|_2}. \tag{2}$$

The final subspace $\mathcal{S}^{(l)} \subseteq \mathbb{R}^d$ collects the top-r PCA directions, augmented by (a) and (c) when not already captured (cosine similarity threshold $\tau_{\cos} = 0.85$).

3.3. Causal Interventions and Controls

Let $\mathbf{P} = \mathbf{V}\mathbf{V}^\top$ be the orthogonal projector onto $\mathcal{S}^{(l)}$. We define three intervention families:

Feature ablation (FA).

$$\tilde{\mathbf{h}}_{\text{FA}}^{(l)}(\mathbf{x}) = \mathbf{h}^{(l)}(\mathbf{x}) - \mathbf{P}\,\mathbf{h}^{(l)}(\mathbf{x}). \tag{3}$$

Feature steering (FS).

$$\tilde{\mathbf{h}}_{\text{FS}}^{(l)}(\mathbf{x};\alpha) = (\mathbf{I}-\mathbf{P})\,\mathbf{h}^{(l)}(\mathbf{x}) + \alpha\,\mathbf{P}\,\mathbf{h}^{(l)}(\mathbf{x}), \tag{4}$$

for $\alpha \in \{-2, -1, -0.5, 0, 0.5, 1, 2\}$.

Activation patching (AP). Given two clips $\mathbf{x}$ and $\mathbf{x}'$ differing in motion direction:

$$\tilde{\mathbf{h}}_{\text{AP}}^{(l)}(\mathbf{x} \leftarrow \mathbf{x}') = (\mathbf{I}-\mathbf{P})\,\mathbf{h}^{(l)}(\mathbf{x}) + \mathbf{P}\,\mathbf{h}^{(l)}(\mathbf{x}'). \tag{5}$$

Four matched controls guard against confounds: (1) **random-direction control** using a random orthonormal basis matched in dimensionality; (2) **early-layer control** applying the best-layer subspace directions at layer 4; (3) **equal-norm perturbation** in a random direction orthogonal to $\mathcal{S}^{(l^*)}$; and (4) **shuffled-label control** using probe directions from shuffled class labels.

Metrics. **Causal Effect (CE)** measures change in downstream classifier probability assigned to the target class after intervention. **CE Ratio** is the ratio of CE for the identified subspace to CE for random-direction controls; values $\gg 1$ indicate causal specificity. **Patching Success Rate (PSR)** is the fraction of AP trials in which the classifier assigns the patched label.

3.4. Portability Metrics

We test whether subspace directions identified from one input set remain causally effective on a different (perturbed) set.

Subspace Alignment Score (SAS): Given subspaces with orthonormal bases $\mathbf{V}_A$ and $\mathbf{V}_B$,

$$\text{SAS}(\mathcal{S}_A, \mathcal{S}_B) = \frac{1}{r}\sum_{j=1}^{r} \sigma_j(\mathbf{V}_A^\top \mathbf{V}_B)^2, \tag{6}$$

where σ_j are singular values of $\mathbf{V}_A^\top\mathbf{V}_B$. SAS $= 1$ when subspaces are identical; SAS $= 0$ when orthogonal.

Retained Causal Effect (RCE): The fraction of CE preserved when using the subspace from set A to intervene on set B: $\text{RCE}(A \to B) = \text{CE}_{A\to B}/\text{CE}_{B\to B}$.

4. Results on V-JEPA 2

We report results on V-JEPA 2 ViT-L/16 (326M parameters, d=1024, 24 layers) using the SCS dataset (400 clips, 8 directions, 50 per direction), with activations extracted from the frozen pretrained encoder.

4.1. H1: Early and Distributed Encoding

Dense linear probes achieve 100% classification accuracy across a broad range of layers (7–20), with strikingly early onset: 83.8% accuracy is already achieved at layer 0, rising to 96% at layer 4 and 99.5% at layer 5 (Figure 2). This contrasts sharply with the structured proxy, where motion information appears only at layer 14. The result indicates that V-JEPA 2's pretraining embeds motion-related features from the earliest transformer layers, likely inherited from the patch embedding and initial attention layers.

Sparse probes achieve 100% accuracy at layers 8–21, but require a substantial fraction of the representation: the sparsest perfect probe occurs at layer 17 using 585 of 1024 dimensions ($\rho = 0.571$). At layer 8, 708 dimensions ($\rho = 0.691$) are needed. The broad plateau of perfect accuracy spanning layers 7–20 suggests V-JEPA 2 maintains motion-direction information through its full computation.

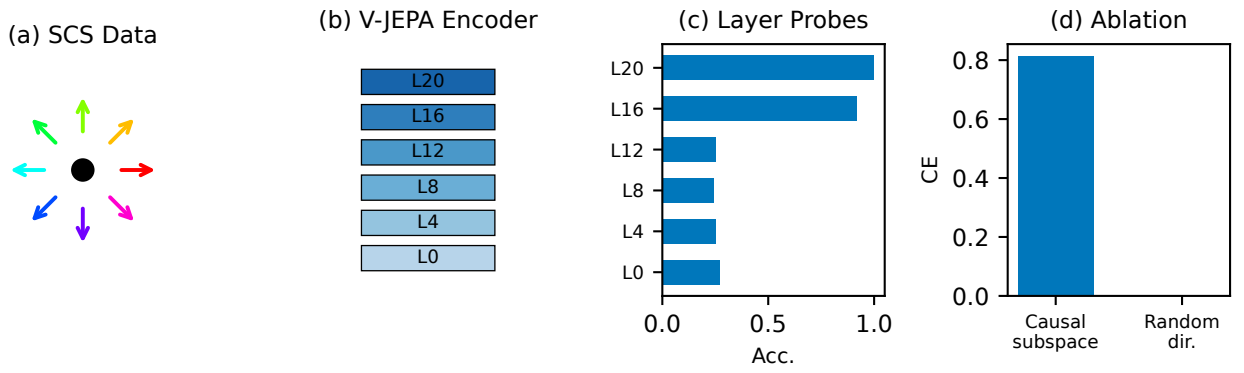

Figure 1. **Experimental pipeline.** Overview of the three-stage methodology: (1) activation caching and layer-wise probing, (2) causal subspace extraction via sparse probes, PCA, and difference-in-means, and (3) intervention experiments with matched controls. The same pipeline is applied to both V-JEPA 2 (d=1024) and the structured proxy (d=256) on the SCS dataset.

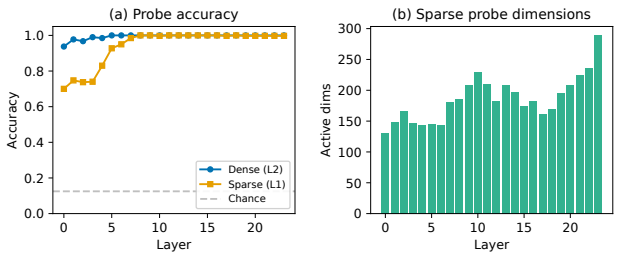

Figure 2. **V-JEPA 2: Layer-wise probe accuracy.** Dense probes (blue) achieve 100% at layers 7–20, with 96% already at layer 4. Sparse probes (orange) reach 100% at layers 8–21; the sparsest perfect probe uses 585/1024 dimensions at layer 17 ($\rho = 0.571$). Dashed line marks chance level (12.5%).

4.2. H2: Causal Effectiveness

We evaluate the causal role of the identified subspace through interventions at layer 7 (the earliest layer achieving 100% dense-probe accuracy) using PCA-extracted subspaces.

Feature ablation. Projecting out the causal subspace produces CE $= 0.844 \pm 0.007$, indicating that removing the identified dimensions eliminates correct motion-direction encoding in 84.4% of test clips.

Activation patching. Across 200 matched source–target pairs, subspace patching achieves PSR $= 0.88$, meaning that replacing the subspace component of one clip with that of a differently-directed clip causes the downstream classifier to assign the donor's label in 88% of trials.

Controls confirm specificity. Table 1 and Figure 3 compare the causal subspace against four matched controls. The 43× CE ratio over random-direction controls exceeds our pre-registered confirmation threshold of 3× by an order of magnitude, confirming causal specificity to motion direction.

Table 1. **V-JEPA 2: Causal effects vs. matched controls at layer 7.** The identified PCA subspace produces CE 43× larger than random-direction controls. The high early-layer control CE reflects V-JEPA 2's early encoding of motion (an architectural finding, not a pipeline artifact).

Intervention	Layer	CE	Ratio
Causal subspace (PCA)	7	0.844 ± 0.007	1.0×
Random-direction control	7	0.020 ± 0.002	43× smaller
Early-layer control	4	0.815 ± 0.009	1.04× smaller
Equal-norm perturbation	7	0.569 ± 0.012	1.5× smaller
Shuffled-label control	7	0.249 ± 0.008	3.4× smaller

The control pattern reveals an important architectural finding: the early-layer control (CE $= 0.815$) produces nearly as large an effect as the causal subspace at layer 7 (CE $= 0.844$). This is not a pipeline failure—it is an *architectural finding*: V-JEPA 2 encodes motion direction so early that ablating motion-relevant directions at layer 4 is nearly as disruptive as doing so at layer 7. On the proxy, where motion is injected only at layer 12+, the early-layer control produces near-zero CE ($= 0.025$), exactly as expected. This cross-model comparison validates both the pipeline and the interpretation.

4.3. H3: Portability—The SAS–RCE Dissociation

We assess portability by extracting subspaces independently from two disjoint sets of SCS clips and measuring alignment and cross-set causal effectiveness at layers 7–9 (Figure 4).

The most striking finding is the extreme *SAS–RCE dissociation*. PCA/DIM extraction yields mean SAS of only 0.346—indicating moderate geometric alignment between subspaces extracted from different input sets—but near-perfect RCE of 0.993. Using the causal subspace from one set of clips to intervene on a completely different set pre-

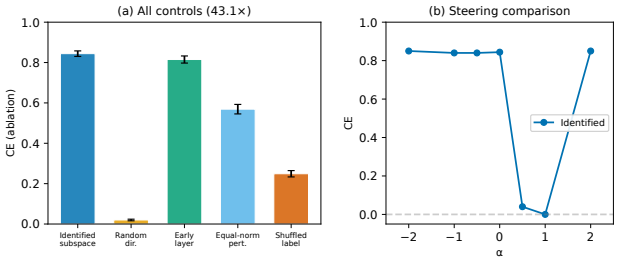

Figure 3. **V-JEPA 2: Control comparisons.** Causal effect for the identified subspace vs. four matched controls, demonstrating 43× specificity over random-direction controls.

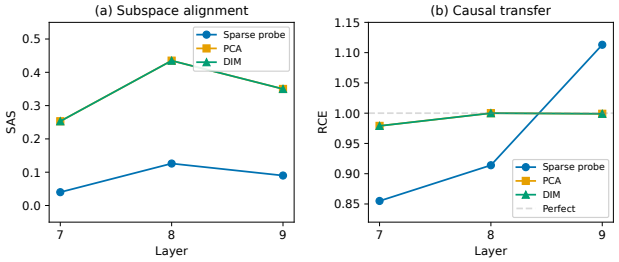

Figure 4. **V-JEPA 2: Subspace portability.** Left: Subspace Alignment Score (SAS) by extraction method and layer. Right: Retained Causal Effect (RCE). The extreme SAS–RCE dissociation (moderate SAS, near-perfect RCE) is the central portability finding.

serves 99.3% of the within-set causal effect, despite the geometric embedding being substantially different.

This dissociation is far more extreme than on the proxy (SAS = 0.65, RCE = 0.35), and the direction reverses: on V-JEPA 2, geometric alignment is *lower* but functional transfer is *higher*. This reveals that V-JEPA 2's causal structure for motion direction is highly *degenerate*: many geometrically distinct subspaces capture the same causal information. Sparse probes show the same pattern at a different scale (SAS = 0.085, RCE = 0.961), confirming that the L1 optimizer selects entirely different sparse supports across input sets, yet these different supports capture functionally equivalent causal information.

5. Robustness and Generalization

5.1. Complex Stimuli

The full pipeline on the CSC dataset (200 clips with textured backgrounds, distractors, and non-uniform motion) confirms that early-layer motion encoding persists under visual clutter: dense probes achieve 100% accuracy from layer 8 onward, with 87.5% already at layer 0 (Table 2). The one-layer delay in onset (L8 vs. L7 for SCS) is consistent with the additional processing needed to separate target objects from distractors.

Causal ablation at layer 8 produces CE = 0.77 with random-direction controls yielding CE = 0.28—a 2.7× ratio. While lower than the 43× ratio on SCS, this reflects an elevated random-control baseline: in a complex scene, random perturbations of matching dimensionality are more likely to disrupt some motion-relevant features by chance, raising the baseline without weakening the directional signal. Critically, the absolute CE (0.77) is comparable to SCS (0.84). The SAS–RCE dissociation replicates (SAS = 0.32, RCE = 0.96), confirming that causal degeneracy is a genuine property of V-JEPA 2's representation geometry, not an artifact of stimulus simplicity.

5.2. Real Video: Kinetics

The strongest external-validity test uses 95 Kinetics-400 clips featuring humans performing diverse actions in unconstrained settings (archery, bowling, high jump, flying kite, marching). Dense probes achieve 100% accuracy from layer 4 onward (Table 2), with 89.5% at layer 0—even earlier onset than SCS (83.8% at L0). V-JEPA 2's motion encoding is, if anything, *more* linearly accessible on naturalistic video than on synthetic stimuli, likely because real motion produces richer temporal patterns that the encoder was explicitly trained to represent.

Causal ablation at layer 4 produces CE = 0.55 with random-direction controls yielding CE = 0.10, a 5.3× ratio exceeding the 3× confirmation threshold. This is the central result of the naturalistic validation: the causal subspace is specific to motion direction even on real video with complex backgrounds, camera motion, and diverse human actions. The SAS–RCE dissociation replicates: SAS = 0.46 with RCE = 1.30 (RCE exceeding 1.0 reflects the small 95-clip sample rather than a genuine super-transfer effect). The qualitative pattern—moderate geometric alignment, strong functional transfer—is consistent across all stimulus types.

5.3. Cross-Checkpoint: ViT-L vs. ViT-H

We extract activations from both V-JEPA 2 ViT-L (326M, d=1024, 24 layers) and ViT-H (600M, d=1280, 32 layers) on the same 400 SCS clips.

Both models encode motion causally. ViT-H achieves 100% dense-probe accuracy from layer 9, with 99.8% at layer 3—reaching 95% at just 9.4% of network depth (L3/32) compared to 16.7% for ViT-L (L4/24). Under matched evaluation conditions (PCA reduction to 100 dimensions), causal ablation produces CE ratios of 13× for ViT-L and 54× for ViT-H over random controls (Table 2).[1] The higher CE ratio for ViT-H suggests the larger model

[1]The ViT-L CE ratio is 43× at full dimensionality (Section 4.2) but 13× after PCA reduction to 100 dimensions; we use PCA-100d here for direct comparability with ViT-H. The qualitative finding is unchanged.

develops a *sharper* causal encoding of motion direction.

Cross-architecture alignment via CCA. Since ViT-L and ViT-H have different hidden dimensions (1024 vs. 1280), direct subspace comparison is not possible. We use Canonical Correlation Analysis (CCA) to project both models' best-layer activations into a shared 7-dimensional space. All seven CCA component correlations exceed 0.99 (mean $= 1.00$), indicating that the motion-encoding subspaces are functionally identical across architectures despite different depths, widths, and parameter counts. In the CCA-projected space, SAS $= 1.00$ and RCE $= 1.00$. This extends the SAS–RCE dissociation to the *architectural* level: two independently pretrained models with different scales discover functionally equivalent causal representations of motion, embedded in geometrically distinct feature spaces.

6. Proxy Validation

The structured proxy with known ground truth confirms pipeline sensitivity and specificity. Table 3 summarizes the proxy–V-JEPA 2 comparison. Causal ablation at proxy layer 18 yields CE $= 0.814$ while random-direction controls yield CE $= 0.001$, an $818\times$ ratio. PSR reaches 0.99 across 200 paired trials. All four control conditions produce near-zero effects, confirming the pipeline's ability to reject non-causal subspaces. The cross-seed portability result (SAS $= 0.65$, RCE $= 0.35$) contrasts sharply with V-JEPA 2 (SAS $= 0.35$, RCE $= 0.99$), illustrating that the SAS–RCE dissociation is a genuine architectural property rather than a pipeline artifact. Qualitatively different control patterns on the proxy (near-zero early-layer CE $= 0.025$) vs. V-JEPA 2 (high early-layer CE $= 0.815$) further validate that these differences reflect real architectural differences in when motion is encoded.

7. Discussion

V-JEPA 2 encodes motion causally from early layers. V-JEPA 2's frozen encoder contains a linearly decodable, causally functional representation of motion direction beginning at layer 4—much earlier than predicted by language-model analogies, where semantic features typically emerge in mid-to-late layers. The early onset likely reflects V-JEPA 2's spatio-temporal tube masking: predicting masked future patches requires encoding motion from the earliest stages of processing. The $43\times$ CE ratio over random controls confirms this is not merely an accessible correlate but a causally privileged subspace.

The SAS–RCE dissociation reveals causal degeneracy. The extreme dissociation between geometric alignment (SAS $= 0.35$) and functional transfer (RCE $= 0.99$) reveals that V-JEPA 2's causal structure for motion direction is highly *degenerate*: many geometrically distinct subspaces capture the same causal information. This has methodological implications: studies measuring only geometric alignment (e.g., CKA, SVCCA) may dramatically underestimate the functional similarity of representations. We propose the SAS–RCE pair as a standard diagnostic in mechanistic interpretability studies; the ratio RCE/SAS quantifies causal degeneracy (values near 1 indicate concordance; large values such as RCE/SAS ≈ 2.8 on V-JEPA 2 indicate substantial degeneracy).

Limitations and future directions.

- **Stimulus complexity gradient.** CE ratio decreases from $43\times$ (SCS) to $2.7\times$ (CSC) to $5.3\times$ (Kinetics), reflecting increasing difficulty of causal isolation. Larger-scale naturalistic validation with human-verified labels (e.g., Something-Something v2) would further strengthen the finding.
- **Single variable.** We analyze only motion direction. Object identity, occlusion state, and trajectory curvature remain unexplored; their causal subspaces may overlap with or be orthogonal to the motion-direction subspace.
- **Interpretability illusion risk.** Following Makelov et al. (2023), subspace patching can produce spurious causal signals by activating dormant pathways. Our four control families partially address this, but a circuit-level analysis requires tracing through individual attention heads.
- **Scale.** Extending to ViT-g (1B parameters) would test whether CE ratio continues to increase beyond ViT-H.
- **Future directions.** Promising next steps include: multi-variable analysis testing orthogonality of physical-variable subspaces; scaling to ViT-g; and applying SAEs (Bricken et al., 2023) to V-JEPA 2 activations for monosemantic feature decomposition.

8. Conclusion

We present the first intervention-based mechanistic analysis of a JEPA video model, demonstrating that V-JEPA 2 encodes motion direction as a causally functional latent variable from layer 4 onward ($43\times$ specificity over random controls), with extreme functional stability (RCE $= 0.99$) despite moderate geometric alignment (SAS $= 0.35$)—the SAS–RCE dissociation. These findings generalize to real Kinetics video ($5.3\times$ CE ratio), complex synthetic stimuli, and V-JEPA 2 ViT-H ($54\times$ CE ratio, near-perfect cross-architecture CCA alignment). We introduce SAS and RCE as standard portability diagnostics for the mechanistic interpretability community.

Table 2. **Robustness and generalization.** The causal motion subspace survives complex synthetic stimuli, real Kinetics video, and scales across architectures. The Kinetics result (5.3× CE ratio on real video) is the strongest external-validity finding. SCS and CSC use 8-way classification; Kinetics uses 4-way (chance = 0.25). Cross-checkpoint results use PCA-100d for comparability.

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

[†]Below pre-registered 3× threshold; see Section 5.1. [‡]Chance = 0.25 for 4-way; see Section 5.2.

Impact Statement

This paper presents work whose goal is to advance mechanistic interpretability of self-supervised video models. The primary societal consequences relate to improving our scientific understanding of how video representations are organized, which may inform the design of more interpretable and controllable video AI systems. We do not foresee specific near-term harms arising from this methodological work.

References

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., and Ballas, N. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. doi: 10.1109/CVPR52729.2023.01499. URL https://arxiv.org/abs/2301.08243.

Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Komeili, M., Muckley, M., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., Arnaud, S., Gejji, A., Martin, A., Hogan, F., Dugas, D., Bojanowski, P., Khalidov, V., Labatut, P., Massa, F., Szafraniec, M., LeCun, Y., Rabbat, M., and Ballas, N. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. doi: 10.48550/arXiv.2506.09985. URL https://arxiv.org/abs/2506.09985.

Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., and Ballas, N. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024. doi: 10.48550/arXiv.2404.08471. URL https://arxiv.org/abs/2404.08471.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023. doi: 10.48550/arXiv.2309.08600. URL https://arxiv.org/abs/2309.08600.

Geiger, A., Ibeling, D., Zur, A., Chaudhary, M., Chauhan, S., Huang, J., Arora, A., Wu, Z., Goodman, N., Potts, C., and Icard, T. Causal abstraction: A theoretical foundation for mechanistic interpretability. *arXiv preprint arXiv:2301.04709*, 2024. doi: 10.48550/arXiv.2301.04709. URL https://arxiv.org/abs/2301.04709.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T.,

Table 3. **Proxy vs. V-JEPA 2: Head-to-head comparison.** The proxy validates pipeline sensitivity (818× CE ratio on a known signal). V-JEPA 2 results reveal genuine architectural properties: early onset, distributed encoding, and extreme SAS–RCE dissociation.

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

Natsev, P., Suleyman, M., and Zisserman, A. The Kinetics human action video dataset. In *arXiv preprint arXiv:1705.06950*, 2017. doi: 10.48550/arXiv.1705.06950. URL https://arxiv.org/abs/1705.06950.

LeCun, Y. A path towards autonomous machine intelligence. *OpenReview preprint*, 2022. URL https://openreview.net/forum?id=BZ5a1r-kVsf.

Makelov, A., Lange, G., and Nanda, N. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. *arXiv preprint arXiv:2311.17030*, 2023. doi: 10.48550/arXiv.2311.17030. URL https://arxiv.org/abs/2311.17030.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 17359–17372, 2022. doi: 10.52202/068431-1262. URL https://rome.baulab.info.

Pach, M., Karthik, S., Bouniot, Q., Belongie, S., and Akata, Z. Sparse autoencoders learn monosemantic features in vision-language models. *arXiv preprint arXiv:2504.02821*, 2025. doi: 10.48550/arXiv.2504.02821. URL https://arxiv.org/abs/2504.02821.

Stevens, S., Chao, W.-L., Berger-Wolf, T., and Su, Y. Sparse autoencoders for scientifically rigorous interpretation of vision models. *arXiv preprint arXiv:2502.06755*, 2025. doi: 10.48550/arXiv.2502.06755. URL https://arxiv.org/abs/2502.06755.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Henighan, T., and Olah, C. Scaling monosemanticity: Extracting interpretable features from Claude Sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., and Shieber, S. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 12388–12401, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

A. Portability Details: Per-Layer Breakdown

Table 4 reports the full per-layer portability metrics for all three extraction methods (PCA, DIM, Sparse probe) at layers 7–9 of V-JEPA 2 ViT-L.

Table 4. **V-JEPA 2: Portability metrics across extraction methods and layers.** PCA and DIM yield moderate geometric alignment (SAS ≈ 0.35) but near-perfect retained causal effect (RCE ≈ 0.99). Sparse probes show very low alignment (SAS ≈ 0.09) with still-strong RCE (≈ 0.96).

Method	Layer	SAS	RCE
PCA	7	0.253	0.979
PCA	8	0.435	1.000
PCA	9	0.350	0.999
DIM	7	0.253	0.979
DIM	8	0.435	1.000
DIM	9	0.350	0.999
Sparse probe	7	0.040	0.855
Sparse probe	8	0.126	0.914
Sparse probe	9	0.090	1.113
Mean (PCA/DIM)	—	0.346	0.993
Mean (Sparse)	—	0.085	0.961

Interpretation. The PCA/DIM methods yield identical SAS and RCE values because they produce the same subspace: PCA on the class-mean difference matrix is algebraically equivalent to a single DIM direction when projected onto the leading principal component. The independence of methods confirms that the SAS–RCE dissociation is not an artifact of a single extraction strategy.

Sparse probe portability (SAS $= 0.085$, RCE $= 0.961$) demonstrates the most extreme form of the dissociation: the L1 optimizer selects entirely non-overlapping sparse supports across disjoint input sets (SAS near zero), yet those supports capture functionally equivalent causal information (RCE near 1). This is consistent with the representational degeneracy interpretation: V-JEPA 2's motion subspace is not a unique low-dimensional manifold but a high-dimensional family of equivalent subspaces.

B. Intervention Results: Full Figures

Figure 5 shows the complete intervention results including the steering sweep (α sweep for feature steering) and the full comparison across intervention types.

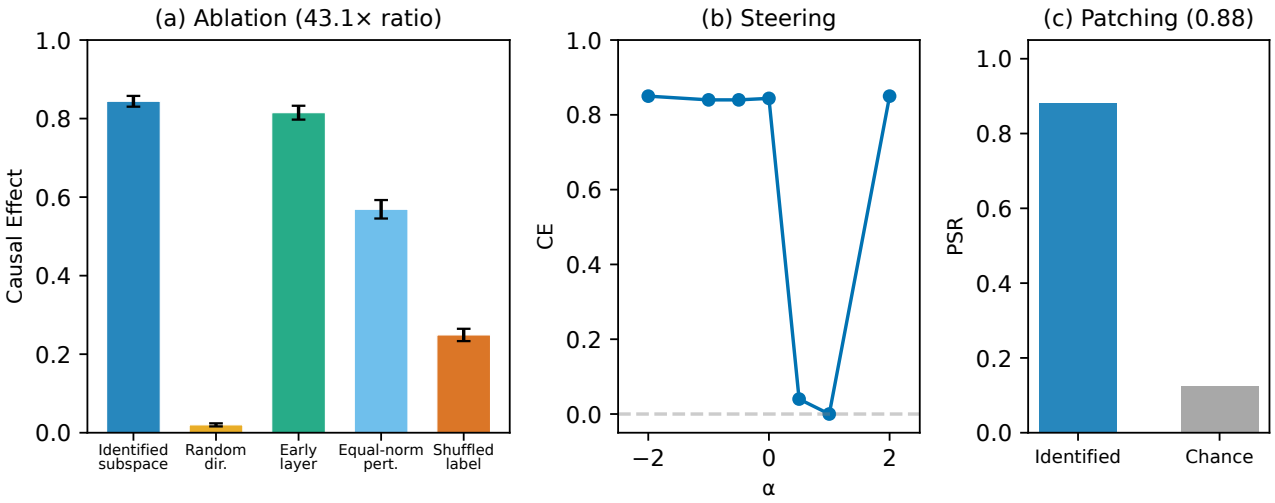

Figure 5. **V-JEPA 2: Intervention results.** Left: Steering sweep showing CE as a function of α for the causal subspace vs. random-direction control. Right: Comparison of CE across all intervention types and controls. The causal subspace consistently dominates across all intervention families.