Reading Frozen World Models Through Action Effects

Tom Jeong
tomjeong0324@gmail.com

July 2026 (scoped draft v0.3)

Abstract

Action-conditioned world models expose both a latent state and a predictor, but most interpretability analyses examine only the former. We ask a complementary question: how does a frozen predictor's latent output vary under controlled actions? We distinguish blockwise snapshot decodability from an off-target decoded-dependence diagnostic, and we analyze centered per-action effect covariances using joint approximate diagonalization followed by action-profile clustering. The procedure observes action identities but uses no state-factor labels during block formation; its outputs are candidate action-selective subspaces, not a certificate of internally modular computation.

In a controlled sprite JEPA, both position factors are decoded from every predefined block at accuracy 1.00, while the primary no-routing checkpoint has zero measured off-target dependence between the encoded factors. After an orthogonal reparameterization, the action-effect analysis recovers factor-selective toy blocks; a separate targeted partial-wall test verifies one y-conditioned right-action prediction that the off-target diagnostic deliberately excludes. On frozen V-JEPA-2-AC, the analysis repeatedly forms small translation-probe-dominant clusters at the supplied finite-difference scales alongside a large residual profile cluster. In a twenty-scene study, blocks are formed transductively from all rollout features and then assigned to Δz using pose supervision on an inner fold; sixteen scenes receive an assignment, without a minimum assignment-quality threshold, and their label-held-out test correlation averages 0.594. A separate confirmatory study fits a two-dimensional response frame independently in each of 48 held-out DROID-seeded scenes from 608 symmetric probe directions. A readout trained on other scenes reaches mean correlation 0.625 with integrated commanded x/z (panel-clustered 95% CI $[0.565, 0.686]$), exceeding a scene-pooled action frame fitted from identical responses by 0.296 ($[0.063, 0.528]$). A corrective antipodal audit leaves the primary result bit-identical and centers the 200-seed permutation null near zero. Fixed readouts otherwise transfer with mixed signs, while cross-scene subspace overlap is elevated but not factor-specific. A second frozen model, DINO-WM on Push-T, yields one profile cluster rather than separated responses. These results support action effects as a useful audit of frozen predictors while leaving causal block function, realized-state coding, and cross-model generality open.

1 Introduction

An action-conditioned world model exposes two related objects: a representation z and a predictor $P(z, a)$. Static probing asks what information is available in z. Action-effect analysis asks how the model's predicted latent output changes when the action changes. Neither question subsumes the other. Snapshot variables can be distributed or redundant while predicted action responses remain selective; conversely, a readable state variable need not be used correctly by the predictor.

Our controlled result illustrates this distinction. In a decoder-free sprite JEPA trained with prediction and Sketched Isotropic Gaussian Regularization (SIGReg) [3], but no routing or explicit

disentanglement objective, nonlinear probes decode both position factors from every predefined latent block at accuracy 1.00. SIGReg discourages collapse by matching random projections of the latent distribution to an isotropic Gaussian. On the same checkpoint, matched pair flips produce zero decoded off-target dependence between those factors. This is not evidence that the latent computation is block diagonal: exact prediction in a factored world implies the observed semantic independence, and even an action-blind identity predictor can pass the off-target diagnostic. The result instead shows that redundant snapshot decodability need not imply off-target predictive dependence.

The predictor also supplies a state-label-free discovery signal. We estimate a centered covariance for each predicted action effect, apply model-specific state preconditioning, jointly approximately diagonalize the covariance family, and cluster rotated coordinates by their action-energy profiles. State labels are reserved for validation, although action channels and probe magnitudes are supplied. In the toy setting, the procedure recovers factor-selective blocks after the entire latent and predictor are co-transformed by an orthogonal rotation. This is a basis-equivariance check, not a structure-destroying null.

We then stress-test the analysis on frozen models we did not train. On V-JEPA-2-AC, two rollout seeds yield several small translation-probe-dominant clusters at the supplied finite-difference scales and a large residual profile cluster. The original pose validation used a row split with trajectory overlap, so we withdraw its pose R^2, coherence verdict, and dependent random-basis significance. A later twenty-scene experiment uses rollout-disjoint supervised scoring after transductive block formation and an inner pose-supervised assignment fold. It supports a narrower result: pose-associated candidate blocks can be found in many scenes, but their fixed readouts are scene-unstable and the recurring subspace family is not factor-specific. On DINO-WM/Push-T, all tested variants return one profile cluster. The positive evidence is therefore one pretrained model family plus controlled toys, not a universal statement about world models.

We also test a simpler query-only object that does not require matching profile clusters across scenes. In each scene, symmetric action probes fit a local two-dimensional response frame without state labels; a global readout is trained on other scenes. Across three disjoint 16-scene panels, the local coordinates transfer to integrated commanded x/z at correlation 0.625. The comparison that survives corrective audit is not a one-draw shuffled null: the local frame exceeds a scene-pooled frame fitted from the same responses by 0.296, with a panel-clustered interval excluding zero. Antipodal closure leaves the primary estimates numerically unchanged while removing the nonzero-mean bias in the original permutation control. This is command-integrator transfer on predictor rollouts, not realized-state recovery or causal identifiability.

Contributions. (1) We separate blockwise snapshot decodability from decoded off-target transition dependence, give the assumptions and counterexamples needed to interpret the latter, and demonstrate the contrast in a controlled JEPA.

(2) We evaluate a state-label-free, action-indexed covariance-JAD procedure that recovers factor-selective candidate blocks after toy latent reparameterization and produces finite-difference-scale-dependent candidate clusters in frozen V-JEPA-2-AC.

(3) We provide an empirical scope audit: transductive block formation followed by rollout-held-out semantic scoring shows scene-unstable fixed readouts and weak factor specificity, while query-only local response frames transfer integrated commands across 48 held-out scenes and DINO-WM supplies a scoped cross-model non-replication.

2 Related work and claim boundary

Post-hoc interpretation of world models. Post-hoc work now probes, intervenes on, and instruments trained world models. Zhang [24] uses Atari-RAM supervision to probe frozen IRIS and DIAMOND representations and intervene along probe-derived directions; Joseph et al. [10] study distributed physical variables in frozen video encoders using probes, geometry, and ablations. Frozen V-JEPA representations have also been analyzed through passive symbol probes [16] and concurrently through causal-state discovery and portability experiments [4]. WorldModelLens supplies a common encode–transition–rollout interface with intervention replay [6]. These results make snapshot analysis a valid complementary object rather than an approach we seek to replace.

Transition diagnostics and fixed-model repairs. ATM is closest in object and protocol: it freezes an encoder and dynamics predictor and uses action-supervised inverse probes to measure action decodability and transfer between real-encoded and model-predicted transitions [7]. Wang et al. [22] test identity, inverse, and composition relations in a pretrained world model before training those relations in. TRM also leaves the world model fixed but learns a supervised trajectory-reachability metric; its small planning-relevant rowspace is obtained from position labels [14]. These works preclude a claim to the first frozen or transition-level world-model diagnostic. Our narrower target is recovery of candidate blocks from a family of action-indexed predictor effects without state-factor labels.

Controllable factors and transition structure. Interventional causal representation learning gives identifiability results under explicit intervention assumptions [1]; BISCUIT jointly learns causal variables and unknown binary interaction variables [15]; and AC-State learns a minimal control-endogenous representation through multi-step inverse models [13]. Action-Controllable Factorization learns independently controllable state variables from pixels [20]. A symmetry-learning lineage jointly learns observation and action representations from transitions [8, 11, 19]. Observed Transition Factorization learns reusable visual transition primitives [17], while FLAM explicitly factorizes latent actions and scene dynamics during training [23]. These works own broad claims about discovering controllable or group-structured factors. We instead audit the basis already present in a frozen explicit-action predictor and make no causal-identifiability claim.

Structured JEPA and sensorimotor training. A concurrent cluster installs structure during training. Causal-JEPA uses object-level masking [18]; SD-JEPA assigns progression and content to fixed subspaces [21]; and Delta-JEPA decodes actions from latent differences [25]. Sensorimotor World Models use inverse dynamics to obtain compact controllable representations [9]. DINO-WM trains a predictor over frozen DINOv2 features but does not recover a post-hoc factor basis [26]. Our inverse-dynamics toy study is therefore an omission-and-repair case study, not a general inverse-dynamics contribution.

3 Measurements and discovery procedure

3.1 Snapshot read selectivity

Let a world state be $s = (s_1, \ldots, s_K)$, let $e = E \circ \text{render}$ map states to the model manifold $\mathcal{M} = e(\mathcal{S})$, and let r_f be a specified probe for factor s_f. For a predefined or discovered coordinate block B_b,

the read matrix is

$$M[b, f] = \text{held-out performance of } r_f \text{ restricted to } B_b.$$

High off-diagonal performance establishes redundant or distributed decodability relative to the chosen blocks and probe class. We use "snapshot redundancy" rather than a coordinate-free claim of entanglement.

3.2 An off-target decoded-dependence diagnostic

Define the decoded semantic prediction

$$\widehat{T}_f(s, a) = r_f(P(e(s), a)).$$

Let μ_g be a matched-pair distribution whose pairs $(s, \tilde{s})$ differ only in factor g, and let ν_f^0 contain actions designated not to move factor f. We estimate

$$J^{\text{off}}[g, f] = \Pr_{\substack{(s,\tilde{s})\sim\mu_g \\ a\sim\nu_f^0}} \left[\widehat{T}_f(s, a) \neq \widehat{T}_f(\tilde{s}, a)\right]. \tag{1}$$

This is a sampled input–output dependence test. It does not measure internal computational coupling, full transition accuracy, or responsiveness to actions. Its diagonal measures retention of the current factor under off-target actions.

Proposition 1 (Exact prediction transfers semantic factorization)**.** *Suppose the world is deterministic, e is injective, and the predictor is exact on $\mathcal{M}$:*

$$P(e(s), a) = e(T(s, a)).$$

If $T_f(s, a) = T_f(s_f, a)$ and $r_f(e(s)) = s_f$, then $\widehat{T}_f(s, a) = T_f(s_f, a)$ and $J^{\text{off}}[g, f] = 0$ for every $g \neq f$.

Proof. Exact prediction gives $r_f(P(e(s), a)) = r_f(e(T(s, a))) = T_f(s_f, a)$. Matched states that differ only in s_g, $g \neq f$, therefore have equal decoded next-f values. □

The converse is false. An action-blind predictor $P_{\text{id}}(e(s), a) = e(s)$ has zero off-diagonal J^{off} and unit diagonal whenever the probes are exact. Equation (1) also excludes target-moving actions, so it can miss a dependence that occurs only while f is being moved—for example, a y-dependent wall that gates a rightward x action. We use J^{off} only as an off-target error-structure diagnostic and test targeted couplings separately.

3.3 Action-indexed effect covariances

For context ξ, probe action $a_\alpha(\xi)$, feature extraction h, and a model-specific reference feature $b_\alpha(\xi)$, define

$$d_\alpha(\xi) = h(P(\xi, a_\alpha(\xi))) - b_\alpha(\xi). \tag{2}$$

The toy uses $b_a(z) = z$; V-JEPA-2-AC uses $b_{\pm i}(z, s) = h(P(z, 0, s))$; and DINO-WM uses the unperturbed prediction under the logged action.

Let L denote the model-specific state preconditioner. The matrices actually supplied to JAD are

$$\widehat{C}_\alpha = \text{Cov}(Ld_\alpha) + 10^{-6}I.$$

They then apply an orthogonal joint approximate diagonalizer [5] and optimize Q to minimize

$$\sum_{\alpha} \frac{\left\| Q\widehat{C}_\alpha Q^\top - \mathrm{diag}(Q\widehat{C}_\alpha Q^\top) \right\|_F^2}{\|\widehat{C}_\alpha\|_F^2}.$$

We optimize an orthogonally parametrized linear layer with Adam and deterministic random restarts. For coordinate i, define

$$e_i = \left([Q\widehat{C}_1 Q^\top]_{ii}, \ldots, [Q\widehat{C}_A Q^\top]_{ii} \right), \qquad p_i = \frac{e_i}{\sum_\alpha e_{i\alpha} + 10^{-9}}.$$

Coordinates whose total energy is below 0.05 times the median are placed in one inert cluster. The remaining singleton clusters are merged greedily: at each step we merge the pair with greatest cosine similarity between their mean normalized profiles, provided that similarity is at least 0.85. The resulting groups are candidate action-selective blocks. Full JAD is performed first; "blocks" arise from centroid-profile clustering afterward.

Model-specific estimands and preprocessing. For the toy, L is full state-covariance whitening with eigenvalues floored at 10^{-3} times the maximum eigenvalue. V-JEPA-2-AC uses the top 96 principal directions of the mean-pooled zero-action predictions, without eigenvalue whitening. DINO-WM analogously uses 64 principal directions of the unperturbed predicted visual features, with common-mode subtraction applied before projection in the corresponding variants. These analyses share a template but not an identical estimator.

Interpretive limits. Writing $C_\alpha = \mathrm{Cov}(d_\alpha)$ for the unpreconditioned population covariance, centered covariance ignores constant effects. If $P(z, a) = z + v_a$, then $\mathrm{Cov}(P(z, a) - z) = 0$ even though the action effect is perfectly selective. A mean-aware extension would instead use

$$G_\alpha = \mathbb{E}[d_\alpha d_\alpha^\top] = C_\alpha + \mathbb{E}[d_\alpha]\mathbb{E}[d_\alpha]^\top$$

and whiten by the pooled effect moment before orthogonal JAD. That corrected estimator is a planned experiment, not the source of the results reported here.

An orthogonal co-transformation $z' = Rz$ gives $C'_\alpha = RC_\alpha R^\top$ and a corresponding solution $Q' = QR^\top$. Scramble survival is therefore expected at population level and tests basis equivariance and numerical stability. The random-Q comparison used in the toy runs is a coordinate baseline, not a confirmatory common-basis null.

4 Controlled experiments

4.1 Sprite world and snapshot–prediction contrast

The controlled world contains cyclic x and y position, color, shape, two distractors, and nine actions including a composite action and a no-op. The primary C2 checkpoint is trained with prediction and SIGReg [3] but no routing or explicit disentanglement objective. Checkpoints are frozen before analysis.

Table 1 shows the primary contrast. Both x and y are decoded from every predefined block at accuracy 1.00. Over the same encoded factors, the estimated off-target matrix is the identity.

The committed exp30 artifact additionally records C5 and C7. In C5, off-target dependence is zero among fully encoded x, y, and color. In C7, including a borderline shape decoder gives

Table 1: Snapshot read accuracy and decoded off-target dependence for the frozen no-routing C2 checkpoint. Color and shape are not reliably encoded and are excluded from this comparison.

Predefined block	x read	y read
G_x	1.00	1.00
G_y	1.00	1.00
G_{color}	1.00	1.00
G_{shape}	1.00	1.00
Free block	1.00	1.00
$J^{\text{off}}[x,x] = J^{\text{off}}[y,y] = 1.00,$		$J^{\text{off}}[x,y] = J^{\text{off}}[y,x] = 0.00$

mean off-diagonal dependence 0.138; the x/y/color submatrix remains zero. We do not claim the five-condition replication stated in an earlier draft because two rows are absent from the authoritative artifact.

4.2 Candidate-block recovery after reparameterization

We co-transform each toy latent and predictor by a fixed random orthogonal matrix before running the discovery procedure. On C10, supervised validation reads x, y, color, and shape at 1.00 from four distinct matched blocks, with color also readable at 0.442 from a fragment block. On never-routed C2, the matched blocks read x at 0.834 and y at 0.971, with other-block reads near chance. The mean matched-minus-best-other separation is 0.711 for C10 and 0.743 for C2, versus 0.00 for one seeded random-Q coordinate baseline. State labels are used for this validation and matching, not for fitting Q or the profile clusters. All 6,000 unlabeled states contribute to whitening, JAD, and profile clustering. The 4,000/2,000 split holds out probe weights and labels, not discovery inputs; the validation is therefore transductive.

4.3 A targeted state-conditioned coupling

A partial wall blocks rightward motion across one boundary only when y lies in a designated interval. This coupling is not visible in $J^{\text{off}}[y,x]$, because Equation (1) excludes actions that move x. We therefore test it directly. At the wall under a rightward action, the predictor keeps x fixed for blocked y values and advances x for free y values; both branches agree with the environment on 1.00 of evaluated cases. This is one controlled example of state-conditioned dependence, not evidence of general contact-physics discovery.

4.4 Secondary omission-and-repair study

The toy objective changes which factors are represented. An anchor-free encoder with per-group variance regularization retains x, y, and color but omits the deliberately difficult shape factor. Oversampling shape-changing actions leaves the self-supervised shape read at chance. Adding a cooperative inverse-dynamics head that predicts the action from a latent pair raises the designated linear shape-slot read from approximately 0.25 to 1.00 in two seeds while preserving the x, y, and color diagonal reads at 1.00.

This result is incremental relative to prior inverse-dynamics work. It also does not produce a globally disentangled representation: x and y leak into the shape slot, nonlinear off-diagonal

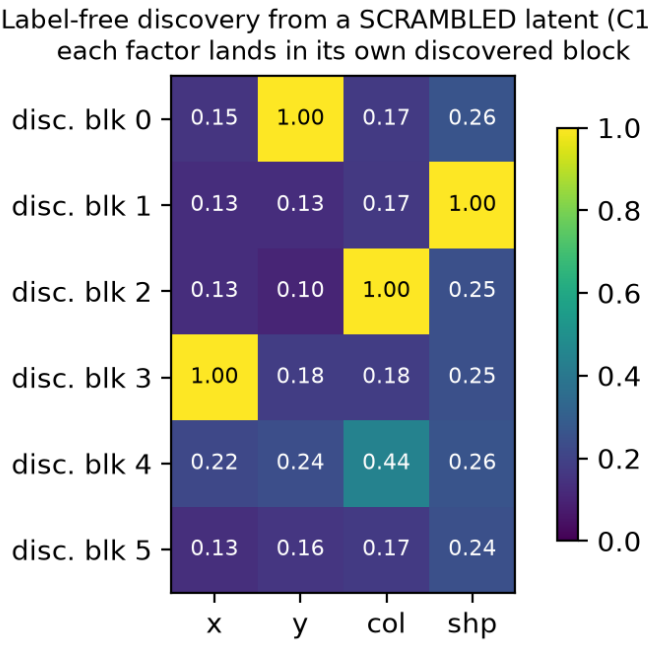

Figure 1: Transductive, state-label-free block formation after an orthogonal reparameterization of frozen C10. Entries are supervised validation accuracies. Four distinct blocks read the four factors at 1.00; a fragment block also carries partial color information. Orthogonal scramble survival is an equivariance check, not a structure-destroying null.

mass remains 0.62–0.64, and shape steering is inconsistent across seeds. We treat it as a controlled omission-and-repair case study.

5 Frozen-model stress tests

5.1 V-JEPA-2-AC action-profile candidates

We freeze the released V-JEPA-2-AC encoder and predictor [2]. The original anchor experiment encodes the first frame from the released two-frame Franka example and generates 384 four-step rollouts under random 7-D end-effector commands, yielding 1,536 predictor states. At each state we evaluate 14 signed probes, two per action channel, against a zero-action prediction; mean-pool the predicted frame tokens; project into the top 96 principal directions of those zero-action predictions; and run covariance JAD and profile clustering.

Across rollout seeds 0 and 2, the procedure returns one large residual profile cluster of 87 and 88 dimensions, respectively, plus several small clusters whose largest raw profile energies occur on the x, y, or z probe pairs. Because profile energies are not normalized for probe magnitude or physical units, this describes the supplied probe configuration and does not establish translation selectivity relative to rotation or grip. The previously reported pose R^2 values are withdrawn: 77 of 461 nominal test rows (16.7%) share a rollout with training rows. The pose-based coherence gate, frame-token pose read, and dependent random-basis significance are therefore not used as confirmatory evidence.

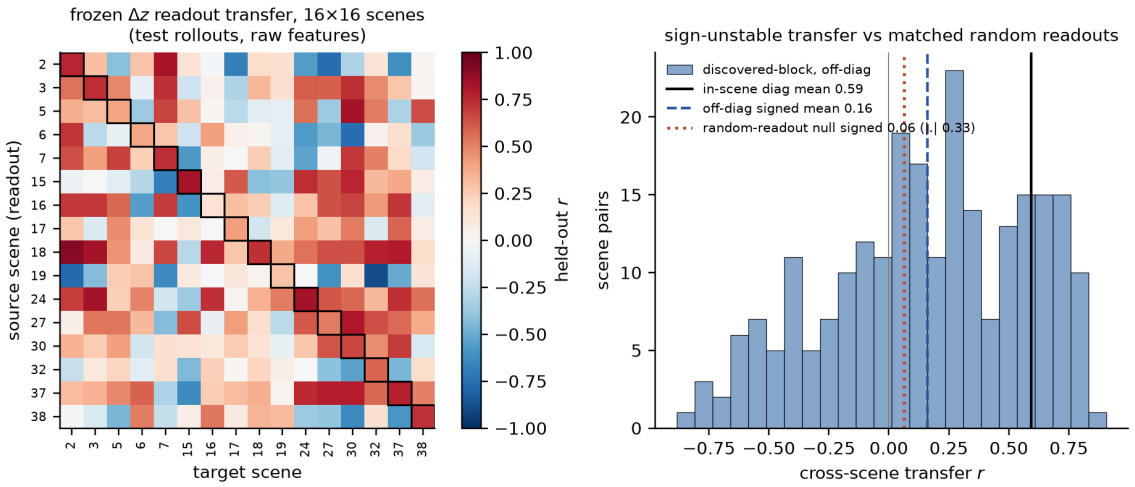

Figure 2: Cross-scene transfer of pose-supervised Δz readouts on label-held-out rollouts after transductive block formation. Fixed weights are partially transferable but vary strongly in sign. Matched random readouts approach the discovered blocks in absolute transfer magnitude.

5.2 Transductive twenty-scene analysis with rollout-held-out scoring

A later experiment repeats per-scene block formation on 20 DROID anchors [12] using 256 three-step rollouts per scene. PCA, effect covariances, JAD, and profile clustering are fit transductively on all rollout features before the rollout masks are applied. Small blocks are then assigned one-to-one to x, y, and z using pose R^2 on an inner validation fold; ridge readouts are fit on the training rollouts and evaluated on label-held-out rollouts. Thus block formation is state-label-free but transductive, while axis assignment and semantic validation are supervised.

Sixteen scenes receive a Δz assignment because they contain an available small block; the assignment procedure has no minimum inner-fold quality threshold. Conditional on those assigned scenes, label-held-out in-scene correlation averages 0.594 and ranges from 0.143 to 0.832. The four unassigned scenes are not included in that mean. This is evidence that action-profile candidates can contain pose-associated command-integrator information across multiple scenes, not an 80% discovery-success rate. It is also not evidence of independently measured robot state: the rollout target is obtained by integrating the commanded actions supplied to the predictor.

5.3 Cross-scene transfer and subspace overlap

Applying each pose-supervised Δz readout to other scenes yields signed off-diagonal mean correlation 0.160, mean absolute correlation 0.371, and 31.7% negative entries. Matched-size random rollout-fit readouts transfer almost as strongly in magnitude (0.330), with signed mean 0.062 and 40.8% negative entries. The random baseline does not repeat the discovered block's assignment-selection step, and transfer cells share source and target scenes, so the sign-coherence difference is descriptive.

Cross-scene Δz–Δz subspace overlap is 0.0354, above a within-span random-subspace reference of 0.0132. This reference matches PCA spans and block dimensions but does not repeat pose-supervised assignment selection. Δz–Δx overlap is nearly as large at 0.0312. Because the pairwise observations share scenes and the primary script does not compute clustered uncertainty for this contrast, we treat the comparison as descriptive. It does not establish a scene-invariant Δz subspace.

Natural DROID scenes vary camera, background, robot configuration, objects, and motion

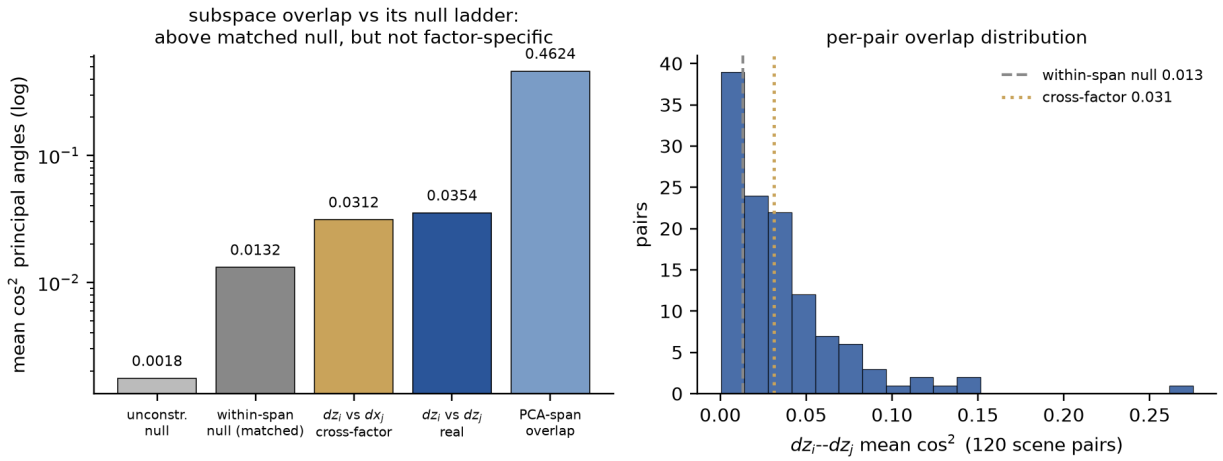

Figure 3: Cross-scene subspace overlap. Pose-associated blocks overlap above a within-span random-subspace reference that matches PCA spans and dimensions but not supervised assignment selection; same-factor overlap barely exceeds the cross-factor control.

distribution jointly. The observed scene instability is consistent with the camera-frame sensitivity reported behaviorally by Assran et al. [2], but it does not identify camera extrinsics as the cause. We therefore make no mechanistic camera-anchoring claim.

5.4 Query-only local response frames

The fixed-readout experiment above asks whether one pose-supervised weight vector transfers across scenes. We separately ask whether each frozen scene can be calibrated without state labels by querying the predictor. For each scene, 608 symmetric action directions fit a two-dimensional local action-response Jacobian on training rollouts. Baseline rollout differences are pulled back through that frame; only the final readout from the two local coordinates to integrated commanded x/z uses labels from other scenes. Held-scene fitting uses known actions and predictor responses but no matched cross-scene state rows.

The confirmatory design contains three scene-disjoint panels of 16 previously unopened DROID-seeded scenes and five action/query seeds per panel. Mean held-scene correlation is 0.625; the three panel means are 0.605, 0.619, and 0.652, giving a panel-clustered 95% interval $[0.565, 0.686]$. A scene-pooled two-dimensional action frame fitted from the identical response rows reaches 0.330; the paired local-minus-pooled gain is 0.296 with panel-clustered interval $[0.063, 0.528]$.

The preregistered shuffled-action gate was invalid as implemented. The original four-direction design has nonzero first moment, so an intercept-free shuffled fit has a nonzero rank-one expectation. We therefore perform a corrective audit using the zero-cost antipodal closure: every direction–response row (u, d) is augmented by $(-u, -d)$, requiring no new model calls. Closure changes the true-pairing Jacobians by at most 9.33×10^{-15} and leaves all primary transfer scores bit-identical. Across 200 permutations for each of the 15 primary caches, the corrected null mean is -0.00028 with panel-clustered interval $[-0.00523, 0.00467]$; the original non-centered null is 0.357. A single corrected draw still exceeds 0.10 in 16.1% of draws, so we report the null distribution rather than reinterpret a one-draw threshold.

This result remains narrow. A pairing-free rank-one response level near 0.357 shows that substantial transfer does not require resolving the full two-dimensional action assignment. More

importantly, the target is integrated commanded displacement generated on model rollouts, not independently measured robot motion. The local-frame result therefore establishes query-efficient command-coordinate calibration for a frozen predictor, not realized-state coding, physical scene alignment, or a globally flat gauge.

5.5 DINO-WM: a scoped non-replication

We apply the same high-level template to frozen DINO-WM on Push-T [26], using perturbations around logged actions and only the predicted visual feature channels. The v2 variants use logged-action and logged-proprio conditioning, optional across-probe common-mode subtraction, and above- versus below-median pusher–block-distance strata. These strata are not verified free-space and contact labels. Every tested variant yields one profile cluster containing all 64 retained PCA dimensions. In an exploratory 70/30 row split, this full retained space decodes agent and block position strongly; that scoring split is not episode-disjoint. The result is therefore an undivided profile response under the present pipeline, not a verified absence of x/y factor structure.

This is a failure of the present feature and intervention pipeline to separate the responses, not evidence that DINO-WM lacks factor structure. Push-T has two physical action dimensions, the current analysis mean-pools tokens, and the estimator does not use pooled-effect whitening. The earlier claim that its low JAD residual is automatically trivial under two actions does not apply to the estimator actually run.

6 Discussion and limitations

The controlled experiments show that blockwise snapshot decodability and decoded off-target dependence can differ substantially. The distinction is useful, but J^{off} is not a transition-factorization certificate. Exact semantic prediction in a factored world implies the observed off-target result, and an identity predictor can imitate it while ignoring actions. Future work must measure target-moving actions, next-state accuracy, and causal response interventions.

The discovery results are also estimator-specific. Centered covariances miss constant action translations; the toy uses state whitening, while the frozen models use state PCA without whitening. Full JAD followed by profile clustering assumes more than generic block structure. A mean-aware, pooled-effect-whitened estimator and a true matched common-basis null are required before making identifiability or certification claims.

The real-model evidence is one positive pretrained family. V-JEPA rollouts encode integrated commands, not independently measured physical state. Pose labels enter block assignment and readout fitting in the multi-scene analysis. Fixed weights are scene-unstable, but random readouts transfer almost as strongly in magnitude and same-factor subspaces barely exceed cross-factor overlap. Query-only local frames improve over a scene-pooled frame, but a pairing-free rank-one response also transfers substantially; neither result establishes realized-state semantics. The natural-scene design does not isolate camera pose.

We make no steering claim. The tracked exp31a result conflicts with a later prose table, and the later successful visualization controller uses supervised nonlinear factor decoders and a predefined action set rather than the state-label-free candidate blocks. The steering figures and environment-success numbers are omitted pending one authoritative rerun.

The DINO-WM result is a scoped negative. Token-level analysis, mixed action directions, episode-disjoint sampling, and additional external action-rich models are the discriminating next tests. The accompanying research roadmap proposes prospective protocols for these extensions and treats abstention or failure as valid outcomes.

Reproducibility and claim status

The repository contains experiment code, result artifacts, figure PNGs, and an experiment ledger. It does not yet support the earlier absolute claim that every paper number and figure regenerates from a clean clone. Several audit-only statistics lack committed machine-readable outputs; relevant toy checkpoints are local rather than tracked; and only a subset of figures has an artifact-only generator.

For this scoped draft, Appendix A identifies which claims are supported by committed artifacts, which are descriptive, and which have been withdrawn. The paper source compiles with `pdflatex/bibtex`; the repository's existing unit tests cover the older group-recovery utilities rather than the world-model experiment gates. A release tag, pinned external-asset manifest, clean-clone figure entry point, and GPU replay are required before public release.

References

[1] Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 372–407. PMLR, 2023. URL https://proceedings.mlr.press/v202/ahuja23a.html.

[2] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

[3] Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025. URL https://arxiv.org/abs/2511.08544.

[4] Guus Bouwens. Causal state variables in V-JEPA 2 latents: Discovery, intervention, and portability. In *ICML 2026 Workshop on Mechanistic Interpretability*, 2026. URL https://openreview.net/forum?id=CC5xKstwmf. Concurrent workshop paper; metadata verified from OpenReview and author materials.

[5] Jean-François Cardoso and Antoine Souloumiac. Jacobi angles for simultaneous diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 17(1):161–164, 1996. doi: 10.1137/S0895479893259546.

[6] Bhavith Chandra Challagundla, Sanskar Pandey, Param Thakkar, Rishikesh Mallagundla, Yugandhar Reddy Gogireddy, Wenhao Lu, Hindol Roy Choudhury, Shravani Challagundla, Mohamed Deraz Nasr, and Spursh Deshpande. One lens, many worlds: A capability-typed interface for world-model interpretability. *arXiv preprint arXiv:2606.09936*, 2026. URL https://arxiv.org/abs/2606.09936.

[7] Jiaheng Chen. ATM: Action-consistency transfer matrix for diagnosing and improving latent world models. *arXiv preprint arXiv:2606.09028*, 2026. URL https://arxiv.org/abs/2606.09028.

[8] Barthélémy Dang-Nhu, Louis Annabi, and Sylvain Argentieri. Disentangled representation learning through unsupervised symmetry group discovery. *arXiv preprint arXiv:2603.11790*, 2026. URL https://arxiv.org/abs/2603.11790.

[9] Petr Ivashkov, Randall Balestriero, and Bernhard Schölkopf. Sensorimotor world models: Perception for action via inverse dynamics. *arXiv preprint arXiv:2606.20104*, 2026. URL https://arxiv.org/abs/2606.20104.

[10] Sonia Joseph, Quentin Garrido, Randall Balestriero, Matthew Kowal, Thomas Fel, Shahab Bakhtiari, Blake Richards, and Mike Rabbat. Interpreting physics in video world models. *arXiv preprint arXiv:2602.07050*, 2026. URL https://arxiv.org/abs/2602.07050.

[11] Hamza Keurti, Hsiao-Ru Pan, Michel Besserve, Benjamin F. Grewe, and Bernhard Schölkopf. Homomorphism AutoEncoder—learning group structured representations from observed transitions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 16190–16215. PMLR, 2023. URL https://proceedings.mlr.press/v202/keurti23a.html.

[12] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

[13] Alex Lamb, Riashat Islam, Yonathan Efroni, Aniket Rajiv Didolkar, Dipendra Misra, Dylan J. Foster, Lekan P. Molu, Rajan Chari, Akshay Krishnamurthy, and John Langford. Guaranteed discovery of control-endogenous latent states with multi-step inverse models. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=TNocbXm5MZ. AC-State; arXiv:2207.08229.

[14] Liangyu Li, Shengzhi Wang, and Qingwen Liu. Beyond euclidean proximity: Repairing latent world models with horizon-matched trajectory reachability metrics. *arXiv preprint arXiv:2605.22164*, 2026.

[15] Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. BISCUIT: Causal representation learning from binary interactions. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 1263–1273. PMLR, 2023. URL https://proceedings.mlr.press/v216/lippe23a.html.

[16] Hung-Ming Liu. Probing the latent world: Emergent discrete symbols and physical structure in latent representations. *arXiv preprint arXiv:2603.20327*, 2026.

[17] Heejeong Nam, Chandradithya S. Jonnalagadda, Harshit Aggarwal, Eric Xu, and Randall Balestriero. Latent actions from factorized transition effects under agent ambiguity. *arXiv preprint arXiv:2606.30544*, 2026. URL https://arxiv.org/abs/2606.30544.

[18] Heejeong Nam, Quentin Le Lidec, Lucas Maes, Yann LeCun, and Randall Balestriero. Causal-jepa: Learning world models through object-level latent interventions. *arXiv preprint arXiv:2602.11389*, 2026.

[19] Robin Quessard, Thomas D. Barrett, and William R. Clements. Learning disentangled representations and group structure of dynamical environments. In *Advances in Neural Information Processing Systems*, volume 33, pages 19727–19737, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/e449b9317dad920c0dd5ad0a2a2d5e49-Abstract.html.

[20] Rafael Rodriguez-Sanchez, Cameron Allen, and George Konidaris. From pixels to factors: Learning independently controllable state variables for reinforcement learning. *arXiv preprint arXiv:2510.02484*, 2025. URL https://arxiv.org/abs/2510.02484. Submitted to ICLR 2026.

[21] Lucas Thil, Jesse Read, Rim Kaddah, and Guillaume Doquet. Subspace-decomposed jepas: Disentangling progression and content in latent world models. *arXiv preprint arXiv:2605.31111*, 2026.

[22] Zijie Wang, Wei Zhang, Weiming Zhang, Fanqi Zhang, Xiao Tan, Yipeng Qin, and Guanbin Li. World models as group actions. *arXiv preprint arXiv:2605.24578*, 2026. URL https://arxiv.org/abs/2605.24578.

[23] Zizhao Wang, Chang Shi, Jiaheng Hu, Kevin Rohling, Roberto Martín-Martín, Amy Zhang, and Peter Stone. Factored latent action world models. *arXiv preprint arXiv:2602.16229*, 2026. URL https://arxiv.org/abs/2602.16229.

[24] Xinyu Zhang. What do world models learn in RL? probing latent representations in learned environment simulators. *arXiv preprint arXiv:2603.21546*, 2026. URL https://arxiv.org/abs/2603.21546.

[25] Zhenghao Zhang, Yuanxiang Wang, Zhenyu Guan, Yujia Yang, Bingkang Shi, Tianyu Zong, Hongzhu Yi, Guoqing Chao, Xingchen Chen, Tiankun Yang, Chenxi Bao, Tao Yu, Jingjing Zhou, and Jungang Xu. Delta-jepa: Learning action-sensitive world models via latent difference decoding. *arXiv preprint arXiv:2606.31232*, 2026.

[26] Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2411.04983; PMLR v267.

A Claim status and artifact provenance

Claim	Primary artifact	Status	Scope
C2 snapshot/off-target contrast	exp30 result JSON	Retained	Committed artifact contains C2, C5, and C7.
Toy scrambled recovery	exp34 C2/C10 JSONs	Retained	State-label-free fitting; supervised matching; one random-Q baseline.
Contact coupling	exp31 contact JSON	Retained	One checkpoint and one visible partial-wall mechanism.
Inverse-dynamics shape repair	exp29 C10 seed JSONs	Secondary	Linear slot read restored; nonlinear leakage remains.
Original V-JEPA pose R^2 and 0/40 null	exp35 seed JSONs (R^2 only); no committed null artifact	Withdrawn	Row split overlaps rollouts; null output is not committed or wired.
Twenty-scene assignment and transfer	exp38b result JSON	Retained	Transductive blocks; pose-supervised assignment; label-held-out test.
Query-only local-frame transfer	exp53 confirmatory null-audit JSON	Retained	48 held-out scenes; command-integrator target; corrective post-hoc null audit.
DINO-WM one-block result	exp39 v2 JSON	Retained	Scoped to state-PCA, centered covariance, mean-pooled visual features.
Steering/environment success	exp31a result JSON	Withdrawn	Tracked JSON conflicts with later prose; successful variant is supervised.

B Implementation details

Toy read and dependence matrices. The read matrix uses independently sampled training and test states and nonlinear probes restricted to predefined blocks. The off-target matrix uses 6,000

matched pairs per source factor. Factors are included only when whole-trace probe accuracy exceeds chance by 0.3. This threshold identifies decodable factors; it does not calibrate the disagreement probability in J^{off}.

Toy discovery. The toy analysis uses 6,000 sampled states, a 20-dimensional latent, state-covariance whitening with an eigenvalue floor, three JAD restarts, cosine-profile clustering at threshold 0.85, and 4,000/2,000 labeled train/test states for post-hoc validation. All 6,000 states enter whitening, JAD, and clustering; only the probe weights and labels are held out.

V-JEPA anchor analysis. The predictor uses 384 four-step random-action rollouts and 14 signed probes over the seven action dimensions. Mean-pooled zero-action predictions define the top-96 PCA span. The original 70/30 row split is retained only as an exploratory artifact and is not used for the main semantic claim.

Multi-scene analysis. Each scene uses 256 three-step rollouts, or 768 states. All states enter unlabeled PCA, JAD, and clustering. Within the supervised stage, assignment-fit and assignment-validation folds partition the readout-training rollouts; the readout-test rollouts are disjoint. Candidate blocks are restricted to at most 20 coordinates. Axis assignment is greedy and one-to-one based on inner-fold pose R^2.

Local response-frame confirmation. Each confirmatory scene uses 64 four-step rollouts and two finite-difference doses. Thirty-eight rollout groups fit the frame, twelve are validation groups, and fourteen are held-out test groups. The primary frame uses four x/z directions and their symmetric predictor responses in a fixed 128-dimensional token sketch. Panels are scene-disjoint and contain 16 scenes each; the five compound seeds vary rollout actions and nested query order. Corrective null inference uses antipodal closure and 200 action-assignment permutations per primary cache. The panel-level t interval has three independent panel means and two degrees of freedom. Full response payloads, query schedules, sketch matrices, labels, and manifests are retained in the audit artifacts.

DINO-WM analysis. The v2 experiment uses 400 Push-T validation windows and four signed probes around logged normalized actions. Common-mode variants are stratified into 200 above-median and 200 below-median pusher–block-distance windows; these are proximity strata, not contact labels. Visual channels are mean-pooled before 64-dimensional state PCA. Post-hoc decoding uses a row split and is not episode-disjoint.

C Corrected and withdrawn analyses

V-JEPA semantic scores. The previous draft reported held-out pose R^2 of 0.79, 0.44, and 0.77 and a 0/40 random-basis result. The semantic split is not rollout-disjoint: 77 of 461 test rows share a rollout with training. The separate random-basis audit is post-hoc, does not rerun JAD on null data, depends on a missing intermediate artifact, and is not wired into the primary verdict. These values are withdrawn from the main claims.

Translation versus rotation. Translation probes use magnitude 0.075 and rotation probes use 0.03 in different physical units. Raw response norms therefore do not establish relative sensitivity or rotation blindness. A dose-normalized response curve is required.

Portability percentages and calibration. An earlier draft quoted 98% retention under supervised block refit. That ratio uses a zero-action predictor-step bridge (0.486 versus a 0.497 full-feature reference). Raw real-frame features give 0.397 versus 0.455, or 87.2%. The underlying audit currently exists as a log rather than a committed machine-readable result, so neither percentage is used as a headline.

Scene sign models. The previous draft reported heuristic sign-only statistics without a named machine-readable calculation artifact. Those values are omitted from the scoped claims and do not identify the source of scene misalignment.

Steering. The tracked JSON reports color success/off-target of 0.667/0.333 for the static controller, 0.583/0.417 for the open-loop supervised-decode controller, and 0.250/0.550 for the closed-loop supervised-decode controller. A later 1.00/0.00 prose result has no matching authoritative exp31a artifact and is withdrawn.

D Secondary negative results

Commutator. In an emergent-collision toy world, raw transition commutators are dominated by generic predictor geometry and position. Controls reduce the apparent contact-localization signal toward chance. This is a powered negative for the tested commutator diagnostic, not a claim that the model lacks collision information.

DROID v1 power failure. The first real-frame portability evaluation used short frame gaps and heterogeneous episodes. Even supervised refit baselines had little signal, so the negative was void for power and motivated the later static-filtered, longer-gap protocol.

DINO-WM. The one-block result remains under all saved v2 variants. Because the current design has two physical action dimensions and mean-pooled features, it scopes the measurement pipeline at least as much as the model.

E Research extensions

The accompanying `paper/RESEARCH_ROADMAP.md` specifies four prospective tracks: active mixed-direction probe selection with calibrated abstention; causal block-only and block-ablation planning interventions; state-conditioned local response atlases; and query-only cross-scene gauge alignment. These protocols are not preregistrations until they are committed, timestamped, and frozen before new runs. The flagship gate requires active identifiability and causal planning specificity on at least two external model/task cells.