LOWER LOSS BY FORGETTING: HOW TRAINABLE TARGETS SUPPRESS INFORMATION IN JEPA WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Joint-embedding predictive architectures (JEPAs) learn world models by predicting future representations and can reduce prediction error by suppressing information already present in their representations. We identify gradients through the learned prediction target as a mechanism for this suppression. Our gradient decomposition shows that, with an optimal predictor and a linear encoder aligned with independent factors, the target contribution attenuates each factor at a rate proportional to its innovation variance. A quadratic covariance penalty yields a closed-form retention threshold and links retained variance to each factor's weight in the linear model's planning distance. Controlled experiments with initially decodable synthetic signals support the predicted ordering, with higher-innovation signals losing decodability earlier. In a simulated pushing task, a model acquires and subsequently suppresses information about block position. A state variable can also remain decodable while having little influence on the latent distance used for planning. On the Cube manipulation task from OGBench, weaker regularisation reduces normalised prediction error by $3.4\times$ while lowering contact-related decodability and planning success. Paired head interventions improve Cube planning, but their benefits depend on the task. These findings identify a mechanism by which predictive training can impair a representation for control, and motivate evaluating information retention and its influence on planning distance alongside prediction accuracy.

1 INTRODUCTION

Lower prediction error can accompany a less useful representation for control. Joint-embedding predictive architectures (JEPAs) learn world models from images by predicting future latent representations (LeCun, 2022; Zhou et al., 2025; Assran et al., 2025; Sobal et al., 2025). In LeWorldModel (LeWM), the encoder and predictor are optimised jointly with a latent prediction loss and SIGReg, a regulariser on the latent distribution (Maes et al., 2026; Balestriero & LeCun, 2025). The planner searches for actions that minimise the Euclidean distance between a predicted latent and an encoded goal image. The learned representation therefore determines both what must be predicted and how the planner compares possible outcomes.

Predictive objectives favour predictable features (Wiskott & Sejnowski, 2002; Richthofer & Wiskott, 2015; Sobal et al., 2022; Littwin et al., 2024; Klindt et al., 2026) and can omit exogenous task-relevant variables (Pendharkar, 2026). State decodability likewise need not imply suitable geometry for planning (Hu et al., 2026; Li et al., 2026; Singh, 2026; Wang et al., 2026). We examine a mechanism connecting these observations: when the encoder also learns the prediction target, it can reduce prediction error by suppressing information that it already represents.

The shared encoder supplies both the predictor's context and its future target. We separate these gradient contributions at an optimal predictor (Fig. 1). For an aligned linear encoder with independent factors, the target contribution attenuates each factor in proportion to its innovation variance, which measures the variation that history and actions cannot predict. In this setting, a quadratic covariance penalty yields a retention threshold and relates each factor's retained variance to its weight in the planning distance.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

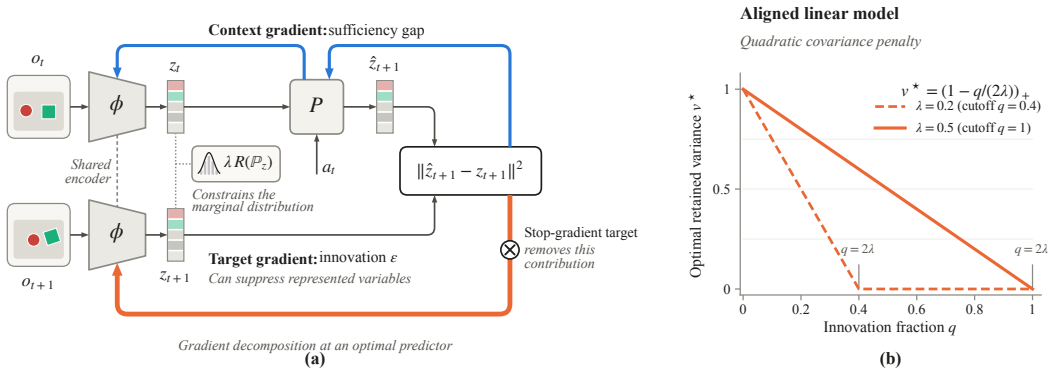

Figure 1: **How target gradients can suppress represented variables.** (a) At an optimal predictor, the context and target gradient contributions involve the predictive sufficiency gap and latent innovation, respectively. A stop-gradient target removes the latter contribution; the regulariser constrains the marginal latent distribution. (b) The aligned linear model with a quadratic covariance penalty has optimal retained variance $v^\star = (1 - q/(2\lambda))_+$ and cutoff $q = 2\lambda$ (Proposition 5); the curves show $\lambda = 0.2$ and 0.5. Proposition 3 separately considers additive independent signals: when the signal-free representation minimises the regulariser, positive innovation incurs an additional prediction cost. The neural retention boundaries in Figure 2 are empirical and need not follow the linear cutoff.

Controlled neural experiments test whether this suppression occurs during learning. We append signals with controlled innovation to fixed visual features and verify that they are decodable before training. Their subsequent loss of decodability therefore distinguishes suppression from a failure to acquire them. Higher-innovation signals lose decodability earlier with target gradients, while stop-gradient targets preserve them in the tested settings. These results support the predicted ordering, although the observed retention boundaries depend on the encoder and need not follow the exact theoretical threshold.

The consequence for control also depends on how retained state affects the planning distance. In a simulated pushing task, a controlled pusher moves a block through contact in the presence of disturbances. A model trained from images acquires and subsequently suppresses information about block position. The same setting also shows that block position can remain decodable while block motion has little influence on the latent distance used for planning. Recovering a variable with a probe therefore does not establish that the planner responds to it appropriately.

On the Cube manipulation task from OGBench (Park et al., 2025), reducing regularisation lowers normalised prediction error by $3.4\times$ while reducing contact-related decodability and planning success. Paired head interventions on fixed image backbones improve Cube planning, but their benefits depend on the task. These findings motivate assessing prediction accuracy, retained state information and the distance used by the planner together.

We make three contributions:

- **A target-gradient mechanism for information suppression** (§2), with explicit conditions distinguishing suppression of represented variables from the absence of an incentive to acquire them, and a linear analysis connecting innovation, retained variance and latent distance.
- **Controlled neural evidence for the mechanism** (§3), using independently varied signal innovation and a simulated pushing environment to examine suppression during training and its effect on distance sensitivity.
- **An evaluation of its implications for planning** (§4), including benchmark loss–retention trade-offs, paired objective interventions on fixed backbones, and settings in which retention does not explain planning performance.

Prospectively recorded predictions, protocol amendments, negative results and exploratory analyses are documented in Appendix D.

2 How prediction training changes the representation

Setting. An agent observes o_t, takes action a_t, and trains an encoder $z_t = \phi_\theta(o_t) \in \mathbb{R}^D$ and an action-conditioned predictor P with the population objective

$$\mathcal{L}(\theta, P) = \tfrac{1}{D}\,\mathbb{E}\big\|P(c_t) - \phi_\theta(o_{t+1})\big\|^2 \;+\; \lambda\,\mathcal{R}\big(\mathbb{P}_z\big), \qquad c_t = (z_{\le t}, a_{\le t}). \tag{1}$$

We denote the prediction term by L_{pred}. The regulariser depends on the marginal latent distribution $\mathbb{P}_z$, as in SIGReg (Balestriero & LeCun, 2025; Maes et al., 2026) and the variance and covariance terms of VICReg (Bardes et al., 2022). With a differentiable target, the prediction gradient reaches the encoder through both c_t and z_{t+1}; a *detached* (stop-gradient) target removes the latter contribution. Both variants train the encoder. A planner scores action sequences by $\|\hat{z}_{t+H} - z_g\|^2$, where z_g encodes a goal image. The image encoder is the *backbone*; the *head* comprises the projector, predictor (including its output head) and action encoder; the projector output is the *latent* used for planning. Innovation is the residual after conditional-mean prediction from a specified history and actions; its fraction of total variance is $q \in [0, 1]$. For physical variables, we estimate this fraction with a finite linear predictor (Appendix B). Proofs are in Appendix A.

Proposition 1 (Target and context gradients)**.** *At the current encoder, let $P^\star(c_t) = \mathbb{E}[z_{t+1} \mid c_t]$ be differentiable with Jacobian G_t. Define the latent innovation $\varepsilon_{t+1} = z_{t+1} - P^\star(c_t)$ and the* sufficiency gap $\delta_t = \mathbb{E}[z_{t+1} \mid o_{\le t}, a_{\le t}] - P^\star(c_t)$. *Holding this optimal predictor fixed, the encoder gradient is*

$$\nabla_\theta L_{\text{pred}} = \underbrace{\tfrac{2}{D}\,\mathbb{E}\big[J_\theta \phi(o_{t+1})^\top \varepsilon_{t+1}\big]}_{\textit{target path}} \;-\; \underbrace{\tfrac{2}{D}\,\mathbb{E}\big[(G_t J_\theta c_t)^\top \delta_t\big]}_{\textit{context path}}. \tag{2}$$

The context term vanishes when the latent context is predictively sufficient ($\delta_t = 0$). Detaching the target removes the target term.

Online and stop-gradient targets have been studied in self-predictive reinforcement learning (Tang et al., 2023; Ni et al., 2024). Equation 2 identifies the context contribution with the predictive information lost by encoding the observations. For a freely varying encoder, the target contribution moves each target embedding towards the average prediction made for it. In the aligned linear model below, its gradient on a factor's loading σ_i is exactly $2\sigma_i q_i / D$. This yields innovation-dependent attenuation by the prediction term. Shared parameters, a non-optimal predictor and regularisation can alter the net update in neural models; the experiments test the resulting qualitative predictions.

Proposition 2 (Removal and non-acquisition of an independent signal)**.** *Let $o_t = (o_t^s, u_t)$, with u independent of (o^s, a) as processes. Let $z_t = \phi(o_t^s)$ be predictively sufficient and $k(u_t) \in \mathbb{R}^D$ have zero mean, finite second moment and innovation $I_k = \mathbb{E}\|k(u_{t+1}) - \mathbb{E}[k(u_{t+1}) \mid u_{\le t}]\|^2 > 0$. Consider $z_t^\tau = z_t + \tau k(u_t)$. (a) With a differentiable target and an optimal predictor for each τ, $L_{\text{pred}}(\tau) \ge L_{\text{pred}}(0) + \tau^2 I_k / D$, so $\tau = 0$ uniquely minimises the prediction error along this family. (b) If the target is fixed at the u-free z_{t+1}, adding u to the context cannot reduce the optimal prediction error. At any predictively sufficient current latent and optimal predictor, the stop-gradient prediction gradient is zero.*

Thus the objective can penalise a factor already represented, which differs from lacking an incentive to acquire independent unpredictable variation (Pendharkar, 2026). Under sufficiency, detaching the target eliminates this removal gradient. It does not guarantee acquisition or retention during training: context gradients can change an insufficient representation (Tang et al., 2023), and other training effects remain (§3.2).

Proposition 3 (The role of a marginal regulariser)**.** *(a) Encoders inducing the same latent distribution have equal regulariser values. (b) In the setting of Prop. 2, suppose $\mathcal{R}$ is a differentiable function of finitely many smooth moments, with differentiation permitted under the expectation. Then $\frac{d}{d\tau}\mathcal{R}(\mathbb{P}_{z^\tau})\big|_{\tau=0} = 0$. The conditions cover fixed-projection SIGReg and smooth VICReg terms; averaging over projections preserves the result under the regularity conditions in Appendix A. (c) If $\mathbb{P}_z$ already globally minimises $\mathcal{R}$, then $\mathcal{L}(\tau) \ge \mathcal{L}(0) + \tau^2 I_k / D$ for every $\lambda \ge 0$.*

Part (b) gives a local, first-order statement about introducing an independent additive signal. It does not exclude a second-order restoring effect or a re-encoding that replaces other variables. Part (c) additionally assumes that the remaining representation already minimises the regulariser; under this condition every $I_k > 0$ incurs an extra cost. Otherwise, regularisation can offset the innovation cost, as the following model makes explicit.

A linear-Gaussian model. Consider a stationary, zero-mean Gaussian state $x_t \in \mathbb{R}^n$ with $\operatorname{Cov} x_t = I$ and $x_{t+1} = Ax_t + Ba_t + w_t$. The zero-mean Gaussian actions and process noise are independent across time and of one another; at each step (a_t, w_t) is independent of $x_{\le t}$, and $\operatorname{Cov} w_t = Q$. Assume *independent factors*: A and Q are diagonal in a common real orthonormal basis $u_1, \dots, u_n$. Factor i then has innovation fraction $q_i = u_i^\top Q u_i$; stationarity gives $\rho_i^2 + \beta_i + q_i = 1$, where ρ_i is its transition coefficient and β_i its action-induced variance. The encoder is linear, $z = Wx$, with the optimal predictor from (z_t, a_t).

Proposition 4 (Prediction-optimal factors)**.** *Under $WW^\top = I_D$ ($D \le n$), the prediction error is $\frac{1}{D}[\operatorname{tr}(WQW^\top) + \|WA(I - W^\top W)\|_F^2] \ge \frac{1}{D}\sum_{i \le D} q_{(i)}$, where $q_{(1)} \le \cdots \le q_{(n)}$. Rows spanning the D least innovative factors attain the bound. This optimal row space is unique if $D < n$ and $q_{(D)} < q_{(D+1)}$. Within this model, the selection depends on actuation through the stationary innovation fractions and is independent of the goal.*

This recovers the selection rule of predictable feature analysis (Richthofer & Wiskott, 2015; Clark et al., 2019). Without independent factors, the bound need not be attained (Appendix A).

Proposition 5 (Attenuation under a covariance penalty)**.** *Replace the constraint by $\lambda\|\operatorname{Cov} z - I\|_F^2/D$, with $D = n$ and $\lambda > 0$. A minimiser is aligned with the factors and assigns factor i the latent variance*

$$v_i^\star = \left(1 - \frac{q_i}{2\lambda}\right)_+. \tag{3}$$

Consequently, (i) factor i is discarded iff $q_i \ge 2\lambda$; (ii) decreasing λ discards factors in decreasing order of innovation; (iii) a factor with $q_i = 0$ retains unit variance; (iv) whenever $\sum_i v_i^\star > 0$, the variance-normalised prediction error $\sum_i v_i^\star q_i / \sum_i v_i^\star$ is non-decreasing in λ.

Thus weaker regularisation can reduce prediction error by suppressing higher-innovation factors. The ordering extends to separable spectral penalties under the conditions in Appendix A. SIGReg is not separable: Appendix E evaluates its ordering numerically, without transferring the exact cutoff $q_i = 2\lambda$.

Corollary 1 (Sensitivity of the planning distance)**.** *In the aligned solution, expressing x in the factor basis gives $\|\hat{z} - z_g\|^2 = \sum_i v_i^\star(\hat{x}_i - x_{g,i})^2$. Each factor is weighted by its retained variance; a discarded factor cannot affect this cost. More generally, write $\phi(s)$ for the encoding of a rendered state. If it is differentiable at s, then $\|\phi(s + \Delta e_i) - \phi(s)\|^2 = \Delta^2 \|\partial_i \phi(s)\|^2 + o(\Delta^2)$.*

The latter measures local sensitivity to a state factor, which decodability alone does not determine. It differs from the change in cost to a fixed goal, whose first derivative also depends on the goal residual (Appendix A). We measure finite changes directly in §3.2. Together, Prop. 5 and Cor. 1 show how reducing prediction error can attenuate variables in the planning cost. They do not establish that retaining those variables is sufficient for successful planning.

3 Controlled tests of representation suppression

We vary innovation independently of image content, then examine pixel-based learning and planning. Unless specified, *retention* is ridge-probe R^2 on episode-disjoint probe data, measuring linear decodability. Ranges span independently trained heads or models; single-head and post hoc analyses are identified below.

3.1 Independent signals with controlled innovation

We append independent AR(1) signals to fixed encoder features and train a newly initialised head for 40 000 steps with the default LeWM objective, selecting the lowest-validation-loss checkpoint. Conditions differ only in whether prediction gradients pass through the target. The signals have innovation fraction q and are independent of the images (Appendix B), testing the independent-signal setting of Props. 2–3. At initialisation, all signals are decodable ($R^2 = 0.89$–0.98 on LeWM's Cube encoder and 0.75–0.85 on DINOv2).

Target gradients suppress already represented signals. With target gradients, the head retains the constant signal ($R^2 \ge 0.99$ in all three Cube heads), but every signal with $q \ge 0.05$ falls to

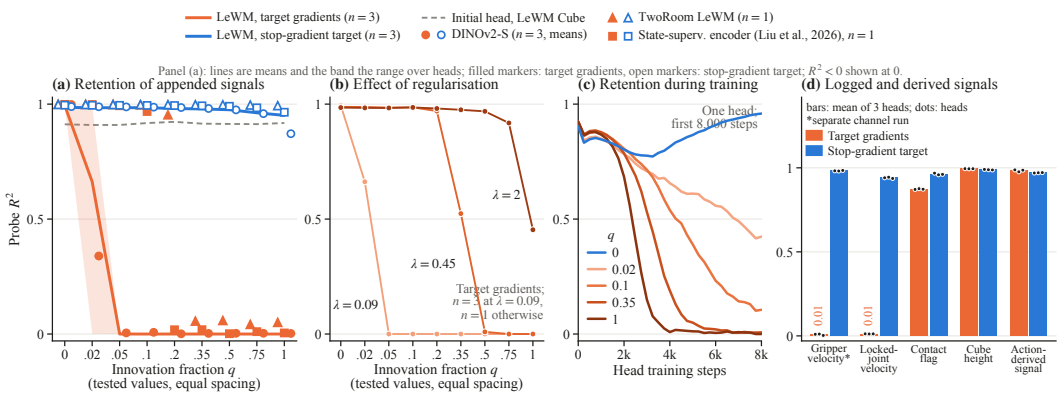

Figure 2: **Independent signals appended to fixed encoder features.** (a) Retention against the innovation fraction q on LeWM's Cube encoder (three heads: mean and range) and other encoders (markers; filled: target gradients, hollow: stop-gradient target); dashed: the initial head on LeWM's Cube encoder. (b) Regularisation strength changes retention (three heads at the default weight, one at each higher weight). (c) Higher-innovation signals lose decodability sooner (one head). (d) Logged robot channels. The action-determined signal does not isolate action conditioning (Appendix B).

$R^2 \leq 0.00$; the $q = 0.02$ signal remains decodable in two of three heads (Fig. 2a). A post hoc two-hidden-layer MLP probe also has $R^2 < 0$ for every signal with $q \geq 0.05$ in every head, corroborating the decline with a nonlinear decoder (Appendix C.1). With a stop-gradient target, all signals remain decodable (0.94–0.99), consistent with Prop. 2(b). The prediction loss is approximately eight times smaller with target gradients (0.0053 versus 0.042); without appended signals the corresponding difference is approximately two-fold. DINOv2-S (Oquab et al., 2024), a generic encoder not trained on these environments, gives the same pattern: $R^2 \leq 0.01$ for $q \geq 0.05$ with target gradients, and $R^2 \geq 0.83$ with a stop-gradient target (three heads each).

Innovation affects the rate and threshold of suppression. The time to halve decodability decreases from 7 500 optimiser steps at $q = 0.02$ to 2 500 at $q = 1$ (Spearman $\rho = -0.97$ and -0.93 in two runs; Fig. 2c). Stronger regularisation preserves signals through $q = 0.2$ at $\lambda = 0.45$ and through $q = 0.75$ at $\lambda = 2$ (one head per weight; Fig. 2b). This agrees with the ordering in Prop. 5, but contradicts the predicted persistence of suppression at both higher weights. The threshold also depends on the encoder: heads on our inverse-dynamics encoder, the state-aligned encoder of Liu et al. (2026), and LeWM's TwoRoom encoder retain signals through $q = 0.05$, 0.1, and 0.2, respectively. Stop-gradient targets retain all signals ($R^2 \geq 0.95$; one head per condition). The common-threshold prediction therefore fails. We establish directional effects, not quantitative agreement with the linear model; neither the encoder dependence nor the regulariser-minimum assumption of Prop. 3(c) is resolved.

Logged robot channels show a similar contrast. The gripper's velocity reading (estimated innovation fraction $q = 0.31$, not decodable from a single frame) falls to $R^2 \leq 0.01$ with target gradients and remains at 0.98 with a stop-gradient target (three heads). Locked-joint jitter behaves similarly; contact and cube height remain decodable in both conditions (Fig. 2d).

3.2 Learning and planning in a simulated pushing task

The MuJoCo scene contains a position-controlled pusher (estimated innovation fraction given actions 0.005–0.008), a free block subject to contact and random forces (0.03–0.05), two visible distractor balls (0.3 and 1), and a per-episode table tint irrelevant to the tasks (Fig. 3a). A third, occluded distractor is excluded (Appendix B). We train a pixel encoder and 16-dimensional latent with the LeWM objective, a stop-gradient target, or an additional inverse-dynamics loss, using three models per condition. Evaluation uses the final checkpoint and an MLP probe, as prospectively specified, and CEM planning with the latent distance to a goal image.

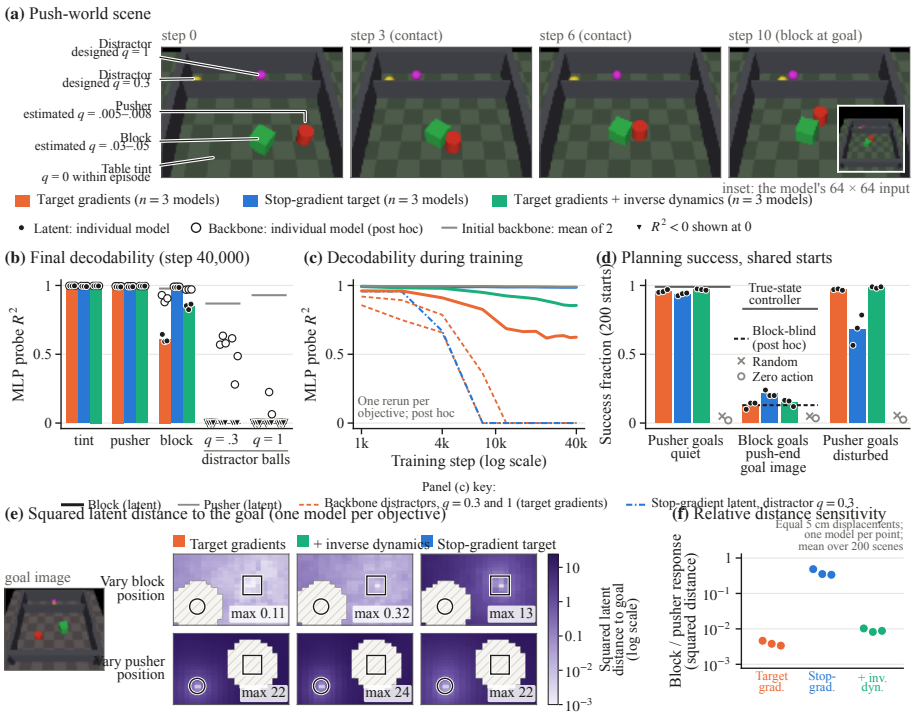

Figure 3: **Representation learning and planning in the push world.** (a) Scene and estimated innovation fractions; inset: the 64×64 input. (b) Final MLP-probe R^2 from the 16-dimensional latent (three models: mean and individual values), the backbone (hollow), and an untrained backbone (grey). (c) Decodability during training (one re-run per objective). (d) Planning on 200 shared starts; block-goal images show the pusher at the end of a push. (e) Goal-cost response to moving only the block (top) or pusher (bottom). (f) Squared latent distance for a 5 cm block displacement relative to an equal pusher displacement. Backbone probes, training trajectories, cost analyses and the block-blind reference are post hoc.

The block is acquired and subsequently suppressed. End-to-end training retains the pusher and tint ($R^2 \geq 0.99$), while block decodability is 0.61 (range 0.59–0.64), above the predicted upper bound of 0.3. A post hoc training trajectory shows that the block is initially acquired: R^2 reaches 0.96 after 1 000 steps and falls to 0.62 by the end, while pusher decodability remains at 1. With inverse dynamics, block decodability falls from 0.99 to 0.86 (one re-run per objective; Fig. 3c). The final end-to-end backbone has higher block decodability (0.91) than the latent. The trajectory establishes temporal suppression; the backbone comparison locates the reduction after the encoder. With a stop-gradient target, the block remains highly decodable (0.99 in every model), contrary to the original prediction.

The distractors reveal a boundary of this mechanism. Both visible balls are decodable from untrained backbones ($R^2 \geq 0.85$), but neither is decodable from any trained latent; end-to-end training also suppresses their backbone decodability. Even with a stop-gradient target, the visible $q = 0.3$ ball reaches $R^2 = 0.96$ during the first 2 000 steps and falls below zero by 8 000 steps (one re-run). Thus acquired content can lose decodability without target gradients, which explain only one source of suppression.

Decodability does not determine the planner's distance. All objectives support pusher goals in a disturbance-free test world (0.93–0.97 success; random actions 0.05). The prespecified block-goal criteria were not met with either goal-image construction, and the predicted advantage of inverse dynamics was not observed (Appendix D). Keeping the pusher at its starting position in the goal image penalises the motion required to push.

Table 1: **Representation and planning on Cube.** Decodability and normalised one-step prediction error span all runs; planning compares one model per row on 300 shared starts, with matched LeWM training splits. Default success spans 58.0–61.7%. Differences use the displayed default and 95% bootstrap intervals over starts. The external model uses state supervision and inverse dynamics without SIGReg; its training split is unknown.

		contact-related R^2		one-step	planning: one model per row	
model	runs	latent	backbone	error (10^{-3})	success (%)	Δ vs. default (pp)
LeWM (default, $\lambda = 0.09$)	3	0.91–0.92	0.94	4.7–5.2	61.7	—
LeWM, weak regulariser ($\lambda = 0.0117$)	3	0.52–0.67	0.77–0.84	1.3–1.5	52.7	-9.0 [-13.7, -4.7]
LeWM + inverse dynamics	1	0.975	0.975	4.8	69.3	$+7.7$ [$+3.0$, $+12.3$]
State alignment + inv. dyn. (Liu et al., 2026)	1	0.979	0.973	6.5	81.3	$+19.7$ [$+15.0$, $+24.3$]

Post hoc goal-cost maps vary one object while holding the goal fixed (Fig. 3e). Separately, controlled scene perturbations measure pairwise latent-distance sensitivity (Fig. 3f): a 5 cm block displacement produces only 0.3–1% of the squared distance produced by an equal pusher displacement for end-to-end and inverse-dynamics models, compared with 34–48% for stop-gradient targets. The inverse-dynamics latent therefore supports block decoding ($R^2 = 0.85$) while responding weakly to block motion in the representation metric, as Cor. 1 allows. With goal images showing the pusher at the end of a push, stop-gradient models improve over a controller that ignores the block and moves the pusher to its goal-image position: 0.21 versus 0.13 success, a paired difference of $+0.08$ $[+0.04, +0.13]$ over 200 starts and three models (end to end: $+0.00$; inverse dynamics: $+0.02$, both with intervals including zero). These diagnostics establish a contribution of the goal cost; dynamics error, search and goal feasibility can also affect success. Under the training disturbance, stop-gradient models achieve lower pusher-goal success (0.68 versus 0.97–0.99), consistent with the drifting block contributing to their cost. These outcomes show that the planning effect of a learned representation depends on the task objective.

4 Benchmark models: retention, interventions and planning

We evaluate LeWM and the state-aligned models of Liu et al. (2026) on Cube, Push-T and TwoRoom. Probe splits are by episode; test episodes are unseen by our Cube world models, but our Push-T suite uses window-level training splits and external training splits are unknown. LeWM's latent-space CEM protocol uses shared starts, with paired bootstrap intervals conditional on trained models (Appendix B). On Cube, *contact-related variables* denotes the gripper opening, contact flag and cube height; their mean probe R^2 summarises decodability. Because gripper opening is commanded, we also report the contact flag separately.

Lower prediction error can accompany reduced retention. A regularisation sweep contains 13 Cube models at eight weights, including three seeds each at the default $\lambda = 0.09$ and weak weight $\lambda = 0.0117$. Reducing the weight lowers normalised one-step prediction error by $3.4\times$, from $(4.7$–$5.2) \times 10^{-3}$ to $(1.3$–$1.5) \times 10^{-3}$, while contact-related decodability falls from 0.91–0.92 to 0.52–0.67 (contact flag alone: 0.83–0.85 to 0.37–0.55). Arm joints (0.97 to 0.99) and planar cube position (0.99 to 0.98) remain decodable, and a nonlinear probe corroborates the contact-related reduction. The weak-model backbones retain $R^2 = 0.77$–0.84, giving backbone–latent gaps of 0.17–0.26, compared with 0.02–0.03 at the default weight. This locates reduced decodability after the backbone without establishing temporal forgetting (Appendix C.4). Mean planning success decreases from 60% to 54% over three seeds and 300 starts. Prediction error increases with regularisation strength (Spearman $\rho = 0.98$), while contact-related decodability saturates above the default weight. These observations support the loss–retention trade-off of Prop. 5(iv), but the estimated innovation of individual Cube variables does not predict which ones are lost.

Head interventions hold the image representation fixed. We retrain only the head on each of the three weak-regulariser backbones. A stop-gradient target improves contact-related decodability relative to a matched end-to-end head on every backbone and for both head initialisations. Planning also improves (Table 2); restricting the fresh-start comparison to the two backbones unseen when the prediction was recorded gives $+4.8$ points $[+2.0, +7.8]$. Inverse-dynamics heads likewise improve

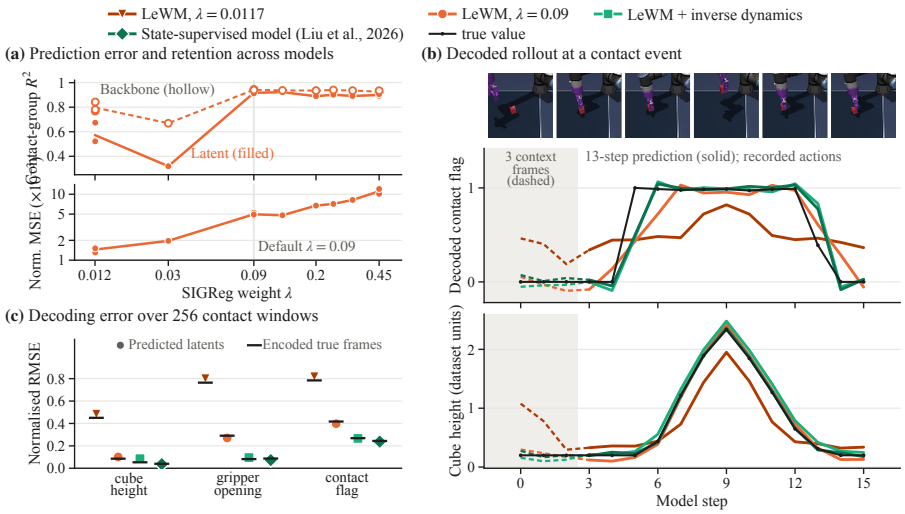

Figure 4: **Contact-related variables in Cube models.** (a) Decodability from the planner's latent (filled) and backbone (hollow), and normalised one-step prediction error, across regulariser weights (13 models; lines connect setting means). (b) Contact and cube height decoded over 13 predicted steps given three context frames and true actions; black: true values. The evaluation window follows a fixed selection rule. (c) Decoding error over 256 evaluation windows, normalised by the factor's standard deviation; ticks: probes on encoded true frames.

Table 2: **Paired head interventions on Cube.** Differences from matched end-to-end heads at $\lambda = 0.0117$, pooled over three fixed backbones. Intervals resample shared starts, preserving the model comparisons. *Batched versus unbatched planning.

head objective	starts	Δ success (pp; 95% interval)
Stop-gradient target	200 fresh	+5.3 [+3.0, +7.8]
Stop-gradient target, second initialisation	100 historical	+6.3 [+2.0, +11.0]
Inverse dynamics*	100 historical	+7.3 [+2.0, +13.0]

decodability on all three backbones (0.84–0.89 versus 0.44–0.77). Their planning difference from stop-gradient heads is 0.0 points $[-4, +4]$ on the same 100 starts. The inverse-dynamics comparison uses batched versus unbatched planning; the measured change from batching alone is -2 points $[-9, +5]$ on these starts.

On a default Cube backbone at $\lambda = 0.0117$, similar contact-related decodability accompanies a +9-point advantage for the stop-gradient head ($p = 0.049$); at $\lambda = 0.09$, their planning performance is similar. On Push-T, the stop-gradient head reaches 44.5% versus 46.5% for the end-to-end head (200 starts), while the original weak-regulariser model exceeds the default model (49.0% versus 40.5%).

One full-model inverse-dynamics intervention. Adding inverse dynamics to full LeWM training on Cube gives contact-related $R^2 = 0.975$ after ten epochs, compared with 0.91–0.92 for the default runs (contact flag: 0.94 versus 0.83–0.85), with equal decodability from backbone and latent. Normalised prediction error remains within the default range (4.8×10^{-3} versus $(4.7\text{–}5.2) \times 10^{-3}$), contrary to the prediction that it would increase. Planning reaches 69.3% on 300 starts, a paired gain of +7.7 points $[+3.0, +12.3]$ over the default model with the same split (Table 1). On 100 additional starts from held-out episodes, success is 64% versus 54% (+10 points $[+3, +18]$). This is one trained model: the intervals describe variation across starts, not training seeds. The gain cannot be attributed specifically to retention, because inverse dynamics also changes the latent geometry and the head experiments show different planning outcomes at similar decodability. No full-model stop-gradient intervention was completed.

State supervision further separates decodability from planning. The model of Liu et al. (2026) combines prediction, simulator-state alignment from consecutive latents, and inverse dynamics, without SIGReg. Its Cube contact-related R^2 is 0.98 and planning success 81.3% (Table 1), but state decodability reflects supervision as well as preservation. A post hoc analysis finds that arm variables explain 56% of latent variance, compared with 7% for LeWM (Appendix C.3). This is consistent with a larger contribution of state variables to latent distance, but does not establish that the resulting metric is appropriate for every task.

The same model improves Push-T success (55% versus 40.5%) despite similar block-angle decodability (0.84 versus 0.81), and TwoRoom success (100% versus 81%), where all models retain agent position ($R^2 \geq 0.99$). These comparisons distinguish decodability from planning geometry, but cannot separate the external model's multiple training objectives (Appendix C.3). Representation of a factor is necessary for the goal distance to respond to it, but decodability alone does not establish an appropriate distance or successful planning.

5 Related work

Predictive objectives favour temporal regularity (Wiskott & Sejnowski, 2002; Richthofer & Wiskott, 2015). Analyses of JEPAs and self-predictive reinforcement learning characterise feature selection, spectral structure and stop-gradient dynamics (Sobal et al., 2022; Littwin et al., 2024; Tang et al., 2023; Ni et al., 2024). LeJEPA admits identifiability guarantees under matched dimensions and isotropic stationary Gaussian dynamics (Klindt et al., 2026). Pendharkar (2026) reports that the evaluated reward-free objectives fail to retain an exogenous task-relevant feature. Multi-step inverse models separate endogenous state from exogenous distractors (Efroni et al., 2022; Lamb et al., 2023; Islam et al., 2023). Recent JEPA variants add inverse dynamics, state supervision or proprioceptive grounding (Ivashkov et al., 2026; Zhang et al., 2026; Liu et al., 2026; Yan et al., 2026).

Prediction accuracy need not track control performance (Lambert et al., 2020; You et al., 2026; Singh, 2026), and state decodability need not ensure suitable Euclidean planning geometry (Hu et al., 2026; Li et al., 2026; Wang et al., 2026). Our contribution is to isolate suppression through trainable targets and relate it to innovation. The spectral selection principle is established; we study how already represented information is lost during training. Our experiments follow initially encoded synthetic signals and a contact-driven object acquired during learning, separating subsequent suppression from non-acquisition. Stop-gradient targets isolate the target contribution, while inverse dynamics tests an auxiliary objective. In the linear model, suppression reduces feature variance and its contribution to the planning cost. Experiments measure the corresponding suppression and cost sensitivity, identifying one mechanism that can impair planning geometry. Appendix F gives the extended comparison.

6 Discussion and limitations

Training a predictor and its target jointly can reduce prediction error by suppressing represented variables. Neither prediction error nor probe accuracy alone establishes suitability for planning: a state variable must remain represented and affect the latent distance appropriately.

The exact attenuation rate and retention threshold apply to the independent-factor linear model; neural experiments support qualitative ordering and encoder-dependent boundaries. The regulariser result assumes independent additive factors and, for the global comparison, a remaining representation that minimises the regulariser. Physical variables are dependent, their innovation is estimated with a restricted predictor, and their suppression on Cube does not follow the measured innovation ordering. Gaussian regularisation may also affect non-Gaussian variables such as binary contact.

Detachment removes the target-gradient contribution but does not guarantee retention: a distractor is acquired and subsequently lost even with a stop-gradient target. Retaining more information also need not improve control. The push-world diagnostics identify a contribution of latent distance, without excluding dynamics error, limited search or goal specification. Paired head interventions improve Cube planning but do not transfer consistently to Push-T; the full-model inverse-dynamics result uses one run and cannot attribute improvement uniquely to retention. These findings motivate joint evaluation of prediction, recoverable state information and the distance used by the planner.

AI USE STATEMENT

Generative AI tools were used in this project. An Anthropic Claude coding agent was used to implement some of the analysis and training code, assist with experiments, check references against primary sources, and proofread and revise the manuscript. It also assisted the authors with preparing the layout of the pipeline figure and arranging the visualizations in the figures. The LLM-generated code was verified and tested for correctness by the authors. We take responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The anonymised supplementary repository provides training and analysis code, logged experimental outputs, and commands for reproducing the reported measurements, tables and figures. It includes the prediction record with timestamps and a record of training and evaluation runs, including unsuccessful runs. Appendix B describes the training procedures, data splits, model-selection rules, probes, innovation estimates and planning evaluations; Appendix A gives the assumptions and proofs; and Appendix D documents prospective and exploratory analyses. The benchmark data and base models use public releases of LeWorldModel, DINOv2 and Liu et al. (2026); the pushing environment is generated by the supplied code. The models trained for this study will be released.

REFERENCES

Guo An, Zijing Wu, Honghua Dong, Yuhao Yan, Zixuan Gui, Haochong Chen, Shanzhao Ruan, Xiang Wang, Yurong Ling, and Qi Tian. Diagnosing JEPA world models with action-conditioned predictive consistency. arXiv:2608.12939, 2026.

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

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. arXiv:2511.08544, 2025.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

David G. Clark, Jesse A. Livezey, and Kristofer E. Bouchard. Unsupervised discovery of temporal structure in noisy data with dynamical components analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Felix Creutzig, Amir Globerson, and Naftali Tishby. Past-future information bottleneck in dynamical systems. *Physical Review E*, 79(4):041925, 2009.

Yonathan Efroni, Dipendra Misra, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Provably filtering exogenous distractors using multistep inverse dynamics. In *International Conference on Learning Representations (ICLR)*, 2022.

Amir-massoud Farahmand, André Barreto, and Daniel Nikovski. Value-aware loss function for model-based reinforcement learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54 of *PMLR*, pp. 1486–1494, 2017.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, pp. 10929–10974, 2023.

Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G. Bellemare. DeepMDP: Learning continuous latent space models for representation learning. In *International Conference on Machine Learning (ICML)*, volume 97 of *PMLR*, pp. 2170–2179, 2019.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Christopher Grimm, André Barreto, Satinder Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5541–5552, 2020.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations (ICLR)*, 2024.

Jiaming Hu, Yan Zheng, and Tian Wang. SCALE: State-calibrated latent embeddings for JEPA planning in the right geometry. arXiv:2608.16287, 2026.

Riashat Islam, Manan Tomar, Alex Lamb, Yonathan Efroni, Hongyu Zang, Aniket Didolkar, Dipendra Misra, Xin Li, Harm van Seijen, Remi Tachet des Combes, and John Langford. Agent-controller representations: Principled offline RL with rich exogenous information. In *International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, 2023.

Petr Ivashkov, Randall Balestriero, and Bernhard Schölkopf. Sensorimotor world models: Perception for action via inverse dynamics. arXiv:2606.20104, 2026.

Khimya Khetarpal, Zhaohan Daniel Guo, Bernardo Ávila Pires, Yunhao Tang, Clare Lyle, Mark Rowland, Nicolas Heess, Diana Borsa, Arthur Guez, and Will Dabney. A unifying framework for action-conditional self-predictive reinforcement learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 258 of *PMLR*, pp. 181–189, 2025.

David Klindt, Yann LeCun, and Randall Balestriero. When does LeJEPA learn a world model? arXiv:2605.26379, 2026.

Alex Lamb, Riashat Islam, Yonathan Efroni, Aniket Didolkar, Dipendra Misra, Dylan Foster, Lekan 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.

Nathan Lambert, Brandon Amos, Omry Yadan, and Roberto Calandra. Objective mismatch in model-based reinforcement learning. In *Learning for Dynamics and Control (L4DC)*, volume 120 of *PMLR*, pp. 761–770, 2020.

Yann LeCun. A path towards autonomous machine intelligence, version 0.9.2. OpenReview, `https://openreview.net/forum?id=BZ5a1r-kVsf`, 2022.

Liangyu Li, Shengzhi Wang, Libin Qiu, Mingliang Xiong, and Qingwen Liu. World model control by trajectory reachability metrics. arXiv:2605.22164, 2026.

Etai Littwin, Omid Saremi, Madhu Advani, Vimal Thilak, Preetum Nakkiran, Chen Huang, and Joshua Susskind. How JEPA avoids noisy features: The implicit bias of deep linear self distillation networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Muyuan Liu, Yue Huang, Zheng Liang, and Xiang Gao. Toward physically grounded JEPA world models for goal-conditioned robotic planning. arXiv:2609.03565, 2026.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. LeWorldModel: Stable end-to-end joint-embedding predictive architecture from pixels. arXiv:2603.19312, 2026.

Tianwei Ni, Benjamin Eysenbach, Erfan Seyedsalehi, Michel Ma, Clement Gehring, Aditya Mahajan, and Pierre-Luc Bacon. Bridging state and history representations: Understanding self-predictive RL. In *International Conference on Learning Representations (ICLR)*, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. OGBench: Benchmarking offline goal-conditioned RL. In *International Conference on Learning Representations (ICLR)*, 2025.

Ayan Pendharkar. Predictive objectives discard exogenous control-relevant features: A controlled mechanistic study. arXiv:2606.30068, 2026.

Stefan Richthofer and Laurenz Wiskott. Predictable feature analysis. In *IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 190–196, 2015.

Max Schwarzer, Ankesh Anand, Rishab Goel, R. Devon Hjelm, Aaron Courville, and Philip Bachman. Data-efficient reinforcement learning with self-predictive representations. In *International Conference on Learning Representations (ICLR)*, 2021.

Joyjeet Singh. The objective is the bottleneck: Latent world models encode what their planners cannot use. arXiv:2608.12959, 2026.

Vlad Sobal, Jyothir S V, Siddhartha Jalagam, Nicolas Carion, Kyunghyun Cho, and Yann LeCun. Joint embedding predictive architectures focus on slow features. NeurIPS 2022 Workshop on Self-Supervised Learning, arXiv:2211.10831, 2022.

Vlad Sobal, Wancong Zhang, Kyunghyun Cho, Randall Balestriero, Tim G. J. Rudner, and Yann LeCun. Learning from reward-free offline data: A case for planning with latent dynamics models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Henning Sprekeler, Tiziano Zito, and Laurenz Wiskott. An extension of slow feature analysis for nonlinear blind source separation. *Journal of Machine Learning Research*, 15:921–947, 2014.

Yunhao Tang, Zhaohan Daniel Guo, Pierre Harvey Richemond, Bernardo Ávila Pires, Yash Chandak, Rémi Munos, Mark Rowland, Mohammad Gheshlaghi Azar, Charline Le Lan, Clare Lyle, András György, Shantanu Thakoor, Will Dabney, Bilal Piot, Daniele Calandriello, and Michal Valko. Understanding self-predictive learning for reinforcement learning. In *International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, 2023.

Claas Voelcker, Tyler Kastner, Igor Gilitschenski, and Amir-massoud Farahmand. When does self-prediction help? understanding auxiliary tasks in reinforcement learning. *Reinforcement Learning Journal (RLC)*, 2024.

Jiawei Wang, Ke Rui, Yushen Zuo, Yichun Feng, and Minglei Li. Decision-metric alignment in latent world models: Diagnostics and action-conditioned objectives for MPC planning. arXiv:2608.18746, 2026.

Ran Wei, Nathan Lambert, Anthony McDonald, Alfredo Garcia, and Roberto Calandra. A unified view on solving objective mismatch in model-based reinforcement learning. *Transactions on Machine Learning Research*, 2024.

Laurenz Wiskott and Terrence J. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715–770, 2002.

Haodong Yan, Jiaguan Zhu, Mingyuan Jia, Ruiqing Yin, Junjie He, Zhide Zhong, Junfeng Li, Jinxuan Lu, Hengtao Li, Tianran Zhang, Jiayi Chen, Wenxuan Song, Wen Chen, Yuxiang Gao, and Haoang Li. Is forward prediction enough? physical state grounding for JEPA world models. arXiv:2608.06799, 2026.

Hanzhe You, Yonggang Zhang, Maohao Ran, Zhiqin Yang, Zhenyuan Zhang, Wei Xue, Jun Song, Xinmei Tian, and Yike Guo. A control theory of predictability in latent world models. arXiv:2607.10362, 2026.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations (ICLR)*, 2021.

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:2606.31232, 2026.

Kai Zhao, Dongliang Nie, Yuchen Lin, Zhehan Luo, Yixiao Gu, Deng-Ping Fan, and Dan Zeng. SubJEPA: Subspace Gaussian regularization for stable end-to-end world models. arXiv:2605.09241, 2026.

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)*, volume 267 of *PMLR*, pp. 79115–79135, 2025.

A PROOFS

Gradient statements assume finite relevant expectations and permit differentiation under the expectation. The first three propositions concern general encoders. The additional assumptions for the linear results are stated below.

Proposition 1. Holding $P = P^\star$ fixed, the chain rule gives

$$\nabla_\theta L_{\mathrm{pred}} = \tfrac{2}{D}\,\mathbb{E}\big[(G_t J_\theta c_t)^\top (P^\star(c_t) - z_{t+1})\big] - \tfrac{2}{D}\,\mathbb{E}\big[J_\theta \phi(o_{t+1})^\top (P^\star(c_t) - z_{t+1})\big].$$

The second term is the target contribution in equation 2. For the first, G_t and $J_\theta c_t$ are measurable with respect to $(o_{\le t}, a_{\le t})$, and

$$\mathbb{E}[\varepsilon_{t+1} \mid o_{\le t}, a_{\le t}] = \mathbb{E}[z_{t+1} \mid o_{\le t}, a_{\le t}] - P^\star(c_t) = \delta_t.$$

The tower property therefore gives the context contribution in equation 2. Although $\mathbb{E}[\varepsilon_{t+1} \mid c_t] = 0$, the parameter Jacobian generally depends on the observations, not only on the latent context. Conditioning on the observations leaves δ_t. Equivalently, writing $\xi_{t+1} = z_{t+1} - \mathbb{E}[z_{t+1} \mid o_{\le t}, a_{\le t}]$ gives $\varepsilon_{t+1} = \xi_{t+1} + \delta_t$: the irreducible residual ξ_{t+1} averages out of the context term. When the envelope theorem applies, the same expression differentiates $\min_P L_{\mathrm{pred}}(\theta, P)$; the fixed-predictor statement does not require that additional conclusion.

To interpret the target contribution, treat the embedding at each observation as freely variable, with functional gradient defined relative to the marginal observation distribution. Its target contribution at o is $\frac{2}{D}\mathbb{E}[\varepsilon_{t+1} \mid o_{t+1} = o]$. A target-only step of size η therefore gives

$$\phi(o) \leftarrow \phi(o) - \tfrac{2\eta}{D}\big(\phi(o) - \mathbb{E}[P^\star(c_t) \mid o_{t+1} = o]\big).$$

This moves the embedding towards the conditional average prediction. It describes one gradient contribution, rather than the net update of a shared neural encoder.

In the independent-factor linear model below, an aligned encoder has loadings σ_i along factors u_i. Its row space is invariant under A, so $WA(I - \Pi_W) = 0$ and the optimal residual is $\varepsilon = Ww$. Consequently $\mathbb{E}[\varepsilon x_{t+1}^\top] = WQ$, the target-gradient component on σ_i is $2\sigma_i q_i / D$, and $\delta = 0$. Detaching the target removes this component. □

Scope. Outside predictive sufficiency, the context gradient can change or discard represented information. For a fixed target it follows the gradient of prediction error; it need not preserve every variable. In linear self-prediction, related analyses connect the learned subspaces to the transition operator (Tang et al., 2023). The proposition does not assert that stop-gradient training prevents all information loss.

Proposition 2. Write $o = (o^s, u)$ and $z = \phi(o^s)$, with the process u independent of (o^s, a). (a) The optimal predictor of z_{t+1}^τ from the latent context cannot improve on the conditional mean given $(o^s_{\le t}, u_{\le t}, a_{\le t})$. By independence, that conditional mean is $\mathbb{E}[z_{t+1} \mid o^s_{\le t}, a_{\le t}] + \tau\mathbb{E}[k(u_{t+1}) \mid u_{\le t}]$. Its two residuals are independent and centred, hence

$$D L_{\mathrm{pred}}(\tau) \ge \mathbb{E}\|z_{t+1} - \mathbb{E}[z_{t+1} \mid o^s_{\le t}, a_{\le t}]\|^2 + \tau^2 I_k.$$

Predictive sufficiency makes the first term $D L_{\mathrm{pred}}(0)$, proving the bound and uniqueness of the minimiser when $I_k > 0$. When $I_k = 0$, the bound remains valid without a strict increase; zero innovation alone does not imply that the perturbed latent remains sufficient or that the prediction error is constant along the family. For an independently represented, unit-variance factor in the aligned linear model, the preceding calculation gives target-gradient component $2\tau q(u)/D$ and zero context contribution.

(b) First hold the target fixed at the u-free z_{t+1} and set $c_t^\tau = (z_{\le t} + \tau k(u_{\le t}), a_{\le t})$. For any predictor P, conditional Jensen gives

$$\mathbb{E}\|P(c_t^\tau) - z_{t+1}\|^2 \ge \mathbb{E}\|\mathbb{E}[P(c_t^\tau) \mid c_t, z_{t+1}] - z_{t+1}\|^2.$$

Independence implies $\mathbb{E}[P(c_t^\tau) \mid c_t, z_{t+1}] = \widetilde{P}(c_t)$, where $\widetilde{P}$ averages $P(c_t^\tau)$ over the independent signal history. It is therefore a predictor from the unperturbed context alone. Minimising over P shows that adding u cannot reduce the optimal error for this fixed target; wherever differentiable,

the derivative at $\tau = 0$ is zero. Second, at any current latent that is predictively sufficient, the stop-gradient update at the optimal predictor contains only the context term of equation 2, which is zero. This is a statement about a gradient with the target held fixed, not a joint objective that changes the target along with τ. □

Acquisition during training. A partially predictable variable can contribute to the prediction of a target that already depends on it; the proposition does not exclude such learning outside the sufficient setting. It establishes neither general acquisition nor general retention under stop-gradient training. In §3.2, the detached-target model retains the block but first represents and later loses the visible $q = 0.3$ ball (Appendix C.2). That observation is outside the sufficient, optimal-predictor case and cannot be attributed to the target-gradient mechanism.

Proposition 3. (a) The regulariser is a function of $\mathbb{P}_{\phi(o)}$; equal distributions give equal values.

(b) Write $\mathcal{R}(\mathbb{P}_z) = F(\mathbb{E}f_1(z), \ldots, \mathbb{E}f_m(z))$, where F is differentiable at the moment vector. A sufficient condition for differentiating each moment is that f_j is continuously differentiable and $\mathbb{E}\sup_{|s|\le s_0} \|\nabla f_j(z + sk(u))\| \, \|k(u)\| < \infty$ for some $s_0 > 0$. Then

$$\frac{d}{d\tau}\mathbb{E}f_j(z + \tau k(u))\bigg|_0 = \mathbb{E}[\nabla f_j(z)^\top k(u)] = \mathbb{E}[\nabla f_j(z)]^\top \mathbb{E}[k(u)] = 0.$$

The chain rule through F proves the claim. For a second-order bound, additionally assume twice differentiable f_j and $\mathbb{E}\sup_{|s|\le s_0} \|\nabla^2 f_j(z + sk(u))\| \, \|k(u)\|^2 < \infty$. Taylor's theorem then gives moment changes of $O(\tau^2)$; a locally Lipschitz F preserves that order. These stronger conditions are needed for the second-order statement, not for the vanishing first derivative.

For fixed projection α and quadrature knots s, SIGReg's Epps–Pulley statistic is a quadratic function of moments $\mathbb{E}\cos(s\alpha^\top z)$ and $\mathbb{E}\sin(s\alpha^\top z)$. The argument applies to a finite set of projections and, under an integrable bound permitting differentiation through the projection expectation, to their average. VICReg's covariance term is polynomial in first and second moments. At positive variance, its variance term is differentiable away from the hinge and locally Lipschitz at the hinge, where the second-order argument applies. A non-smooth functional of the distribution need not satisfy the proposition: for example, the cube root of a centred third moment can vary linearly along the additive family.

(c) If $\mathbb{P}_z$ globally minimises $\mathcal{R}$, then $\mathcal{R}(\mathbb{P}_{z^\tau}) \ge \mathcal{R}(\mathbb{P}_z)$ for every τ. Adding this inequality to Prop. 2(a) proves the total-objective bound for every $\lambda \ge 0$. □

Scope. Part (b) is a local directional-derivative statement; it does not establish stability after a small perturbation. When the total objective is twice differentiable, its curvature determines that stability, and a second-order decrease in the regulariser can offset an increase in prediction error. Part (c) is global along the additive family but assumes the remaining representation already minimises the regulariser. Neither result covers replacing other variables with u. In the linear model, replacing a more innovative factor with a less innovative one can reduce the objective.

Linear model and prediction error. For the remaining results, x_t is stationary Gaussian with $\mathbb{E}x_t = 0$ and $\operatorname{Cov} x_t = I$. It follows $x_{t+1} = Ax_t + Ba_t + w_t$. The zero-mean Gaussian actions and process noise are independent across time and of one another; at each step (a_t, w_t) is independent of $x_{\le t}$. Write $\Sigma_a = \operatorname{Cov} a_t$ and $Q = \operatorname{Cov} w_t$. Stationarity implies

$$AA^\top + B\Sigma_a B^\top + Q = I.$$

When A and Q are diagonal in the common real orthonormal basis $\{u_i\}$, $\rho_i = u_i^\top A u_i$, $\beta_i = u_i^\top B\Sigma_a B^\top u_i$ and $q_i = u_i^\top Q u_i$ satisfy $\rho_i^2 + \beta_i + q_i = 1$. The factor coordinates have unit variance and, conditional on their current values and the action, innovations of variance q_i. Gaussianity makes the optimal conditional-mean predictor linear.

Let $z = Wx$ with $W \in \mathbb{R}^{D\times n}$, allowing zero singular values, and define $\Pi_W = W^\top(WW^\top)^\dagger W$, the orthogonal projector onto the row space of W. The superscript † denotes the Moore–Penrose pseudoinverse. The optimal prediction of $z' = Wx'$ from (z, a) is

$$\hat{z}' = WAW^\top(WW^\top)^\dagger z + WBa = WA\Pi_W x + WBa.$$

Therefore the residual is

$$r = z' - \hat{z}' = WA(I - \Pi_W)x + Ww, \tag{4}$$

a sum of uncorrelated terms, and $\mathbb{E}\|r\|^2 = \mathrm{tr}(WQW^\top) + \|WA(I - \Pi_W)\|_F^2$.

Proposition 4. Under $WW^\top = I_D$, $\Pi_W = W^\top W$ and

$$J(W) = \tfrac{1}{D}\Big[\mathrm{tr}(WQW^\top) + \|WA(I - W^\top W)\|_F^2\Big]. \tag{5}$$

The second term is non-negative. By the Ky Fan minimum principle, $\mathrm{tr}(WQW^\top) \geq \sum_{i \leq D} q_{(i)}$. Under independent factors, choose the row space $S = \mathrm{span}\{u_i : i \in I\}$ for any D factor indices. Then $A(I - \Pi_S) = \sum_{j \notin I} \rho_j u_j u_j^\top$ and $Wu_j = 0$ for $j \notin I$, so $WA(I - \Pi_S) = 0$. Choosing the D smallest-q factors attains the bound. If $D < n$ and $q_{(D)} < q_{(D+1)}$, equality in the trace bound uniquely specifies this row space. For $D = n$, the row space is the whole state space. Actuation enters the selection through the stationary innovation fractions, and no goal appears. □

Without independent factors the bound need not be attained. Take $n = 2$, $D = 1$, $A = r\left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right]$ with $0 < r < 1$, $Q = (1 - r^2)I$ and $B = 0$. Stationarity holds and A commutes with Q, but A has no real eigenvectors. For any unit row $W = w^\top$, $\mathrm{tr}(WQW^\top) = 1 - r^2$ and

$$\|w^\top A(I - ww^\top)\|^2 = \|A^\top w\|^2 - (w^\top Aw)^2 = r^2.$$

Thus $J(W) = 1$, strictly above $1 - r^2$. Commutation alone does not suffice. The second term in equation 5 accounts for dynamics outside the represented subspace. Under independent factors, the selection rule recovers predictable feature analysis (Richthofer & Wiskott, 2015); related self-prediction results characterise learned invariant subspaces (Tang et al., 2023; Voelcker et al., 2024).

Proposition 5. Write $W = V\Sigma U^\top$ with orthogonal V, U and $\Sigma = \mathrm{diag}(\sigma_k)$. Let U_+ contain the columns with $\sigma_k > 0$, so $\Pi_W = U_+U_+^\top$. The objective is

$$\tfrac{1}{D}\Big[\mathrm{tr}(\Sigma^2 U^\top QU) + \|\Sigma U^\top A(I - U_+U_+^\top)\|_F^2 + \lambda \sum_k (\sigma_k^2 - 1)^2\Big].$$

For fixed Σ, von Neumann's trace inequality gives $\mathrm{tr}(\Sigma^2 U^\top QU) \geq \sum_k \sigma_{[k]}^2 q_{(k)}$: larger loadings are paired with smaller innovations. Choosing the columns of U as the common factor basis attains this bound and makes the second term zero. An optimum is therefore factor-aligned. Writing $v_i = \sigma_i^2$, optimisation separates into

$$\min_{v_i \geq 0} \ \tfrac{1}{D} \sum_i \big[v_i q_i + \lambda(v_i - 1)^2\big].$$

For $\lambda > 0$, each scalar objective is strictly convex with unconstrained minimiser $1 - q_i/(2\lambda)$, which gives equation 3. The minimising variance decreases with q_i, consistent with the pairing. This proves (i)–(iii). For distinct q_i, the encoder is unique up to an output rotation.

For (iv), the normalised error is the weighted mean of q_i with weights $v_i^\star$. For $q_i > q_j$ when both are retained, $v_i^\star / v_j^\star = (2\lambda - q_i)/(2\lambda - q_j)$ increases with λ. Newly retained factors also have larger q than those already retained. Increasing λ therefore shifts relative weight towards higher innovation, so the weighted mean is non-decreasing. If every factor is discarded, the normalised error is undefined and the claim does not apply. □

General separable penalty. Replace the covariance penalty by $\lambda \sum_i \psi(v_i)/D$. Assume ψ has a unique global minimum at $v = 1$ and the scalar objectives $f(v; \vartheta) = \vartheta v + \psi(v)$ have minimisers on $v \geq 0$, where $\vartheta = q/\lambda$. For $\vartheta_1 < \vartheta_2$, optimality of minimisers v_1, v_2 at the respective values gives

$$\vartheta_1 v_1 + \psi(v_1) \leq \vartheta_1 v_2 + \psi(v_2), \qquad \vartheta_2 v_2 + \psi(v_2) \leq \vartheta_2 v_1 + \psi(v_1).$$

Adding yields $(\vartheta_2 - \vartheta_1)(v_2 - v_1) \leq 0$, hence $v_2 \leq v_1$. This establishes the attenuation ordering in (ii); for $q = 0$, uniqueness of the penalty minimum gives $v = 1$, as in (iii). If ψ is convex with a right derivative at zero, $v = 0$ is a minimiser iff $\vartheta + \psi'_+(0) \geq 0$. These results need not give the quadratic penalty's cutoff or its normalised-error monotonicity. SIGReg is not a separable spectral penalty; its ordering is assessed numerically in Appendix E.

Corollary 1. In the aligned solution, $z = \sum_i \sigma_i (u_i^\top x) e_i$ up to an output rotation. Orthogonal rotations preserve squared distance, giving $\|\hat{z} - z_g\|^2 = \sum_i v_i (\hat{x}_i - x_{g,i})^2$ in factor coordinates.

For a general encoder, write $\phi(s)$ for the composition of rendering and encoding. Differentiability at s gives $\phi(s + \Delta e_i) - \phi(s) = \Delta\, \partial_i \phi(s) + o(\Delta)$, and squaring gives the stated $\Delta^2 \|\partial_i \phi(s)\|^2 + o(\Delta^2)$. This is a pairwise latent-distance sensitivity. For a fixed goal z_g, the change in its cost is instead

$$\|\phi(s + \Delta e_i) - z_g\|^2 - \|\phi(s) - z_g\|^2 = 2\Delta\, \partial_i \phi(s)^\top (\phi(s) - z_g) + o(\Delta).$$

Thus the same encoder Jacobian governs the local change, but its effect on a fixed-goal cost also depends on the goal residual. Twice differentiability permits an $O(\Delta^2)$ remainder in this last expression. In the aligned linear model, equation 3 and Prop. 5(iv) make both retained variances and normalised error non-decreasing in λ; this explains their joint reduction when regularisation is weakened, without implying a general ranking of planning performance. □

B Experimental details

World models and training procedure. The benchmark models share the LeWorldModel architecture (Maes et al., 2026): a ViT-tiny encoder (patch 14, 224^2 inputs) whose CLS token is mapped by a two-layer batch-norm MLP projector (hidden width 2048) to a 192-dimensional latent; an action-conditioned causal transformer predictor (six layers) with a projection head; frame skip 5 and a context of three frames. Our full-model training minimises next-embedding MSE plus SIGReg (weight λ, default 0.09) with AdamW (learning rate $5{\times}10^{-5}$, weight decay 10^{-3}), batch 128, mixed precision, gradient clipping at 1.0, and the authors' warm-up–cosine schedule defined over 100 epochs and stopped after ten. Our Cube models use an episode-level 90/10 split keyed to the training seed; the original Cube configuration and our Push-T model suite use window-level splits. Thus probe test episodes are independent of world-model training for our Cube models, but not for the Push-T suite. For all models we train, we advance the learning-rate schedule once per optimiser step: the authors' script defines the schedule in optimiser steps (about 1.4 million for Push-T at 100 epochs) but advances it once per epoch, so with the authors' configuration the learning rate is still in the linear warm-up ramp after ten epochs (3.6×10^{-8}, 0.07% of the nominal 5×10^{-5}); stepping per optimiser step gives the schedule the LeWM paper describes (about one epoch of warm-up to 5×10^{-5}, then 4.9×10^{-5} at the epoch-10 stop). Training stops after ten of the 100 scheduled epochs, as in the public training procedure, and we evaluate the epoch-10 checkpoints. The inverse-dynamics variant adds an MLP ($2 \cdot 192 \to 512 \to$ action block) that predicts the action block from two consecutive latents, trained jointly with weight 1 and discarded afterwards; the stop-gradient variant stops prediction gradients into the target embedding. Within each experiment, the compared objectives otherwise use the same optimisation settings.

Head retraining and appended signals. A head is everything after the image encoder: projector, predictor, prediction head and action encoder. We hold a trained encoder fixed, cache its CLS features on the model-step grid (every fifth frame, the only frames a training window touches), and train a freshly initialised head with the procedure above for 40 000 optimiser steps (20 000 on TwoRoom, which has about five times fewer training windows), keeping the weights with the lowest validation loss; for all but two heads (the stop-gradient heads on the state-aligned Cube encoder and on TwoRoom) these weights lie within the last tenth of training. For the appended-signal test the projector receives the CLS token concatenated with the standardised signals (the CLS features have standard deviation ≈ 0.9). The synthetic signals are stationary AR(1) processes on the model-step grid, $x_{k+1} = \sqrt{1-q}\, x_k + \sqrt{q}\, \epsilon_k$, $\epsilon_k \sim \mathcal{N}(0,1)$, with $q \in \{0, 0.02, 0.05, 0.1, 0.2, 0.35, 0.5, 0.75, 1\}$ ($q = 0$ is constant within an episode); their empirical innovation fractions, estimated from their own history, match the design to 10^{-3}. A further signal is fixed by the action block the predictor receives ($u^\top a_{5k:5k+5}$ for a fixed random unit vector u); because the demonstrator's actions are predictable from the frames (ridge R^2 0.42–0.45 from the current CLS token, 0.72 from three past ones), and because the signal is still kept when the predictor receives the actions of other windows, we do not read its retention as evidence about action conditioning. On Cube we also append logged robot channels: the contact flag and cube height (both visible in the images), the velocity of an arm joint that the demonstrations keep locked (position s.d. $< 10^{-3}$ rad; image-invisible jitter) and, in a separate run, the gripper's velocity reading. Three head seeds are used on LeWM's Cube encoder and on DINOv2-S/14 (Oquab et al., 2024) (its 384-dimensional CLS token on LeWM's preprocessed

frames, projector $384 \to 2048 \to 192$, otherwise identical); one head each on the other encoders and at the other regulariser weights. A validity check: the injected head decodes the gripper opening, which is not appended, as well as the plain head on the same backbone (0.94 vs. 0.92). Dense runs log retention every 250 steps for the first 8 000 steps.

Probes and the quantities reported. *Retention* is the R^2 of a ridge probe on standardised features, with the regularisation $\alpha \in \{0.1, 1, 10, 100, 1000\}$ chosen on an inner 80/20 split of the probe's training frames and the probe refitted on all of them, evaluated on frames of held-out episodes with the conventional (test-variance) R^2; features are the 192-dimensional latent used by the planner or the encoder's CLS token before the projector. Probe R^2 measures decodability under the specified decoder and data distribution, not total information content. Probes use 20 000 frames of training episodes and 8 000 of held-out episodes on Cube (10 000 and 4 000 for the model-suite evaluation). The *nonlinear probe* of the pixel worlds (and, post hoc, of the six appended-signal heads on LeWM's Cube encoder, Appendix C.1) is a two-hidden-layer MLP (width 256, GELU) trained with Adam (10^{-3}, 3 000 steps of batch 512) on standardised features and targets, evaluated on held-out episodes in the same way. The *contact-related group* on Cube is the mean R^2 of the gripper opening, the contact flag and the cube height; the *arm* group averages the joint positions, and *cube position* its planar coordinates. The *one-step error* is the mean squared error of the predictor's next-latent prediction from a three-frame context with the true actions, divided by the latent's variance, on held-out windows; rollout error is the same quantity k steps ahead. The *estimated innovation fraction* of a logged or simulated factor is obtained by ridge regression of its next value on its three previous values (and, where stated, the action), as the residual variance over the total variance, on the training data. This restricted regression estimate need not equal the theoretical innovation conditioned on the full context. For synthetic signals, innovation is set by construction. The *backbone–latent gap* is the difference between the group R^2 from the CLS token and from the latent. The *share of variance* a factor group explains is the R^2 of least squares from the standardised factors and their squares to the latent, averaged over latent dimensions weighted by their variance. Factor groups can overlap, so their explained-variance shares need not sum to one. The *participation ratio* is $(\sum_i \sigma_i^2)^2 / \sum_i \sigma_i^4$ of the latent covariance. *Latent-distance sensitivity* in the push world is the mean squared latent distance between the encodings of two rendered scenes that differ only in the block's position (or only in the pusher's) by 5 cm, over 200 random scenes.

Models trained by others. We evaluate the public LeWorldModel models and the public models of Liu et al. (2026), whose Cube and Push-T models we import into the LeWM evaluation. We retain the encoder, projector, predictor, prediction head and action encoder, discard the training-only heads, and check each imported model using its one-step prediction error and a small planning evaluation. Their training objective is next-embedding prediction plus a state-alignment term, which regresses the simulator state from two consecutive latents, plus an inverse-dynamics term, without SIGReg; their planner is LeWM's latent-space CEM, so our planning protocol matches theirs in kind but not in its starts, goal offsets or budgets, and our numbers are not those of their paper. Their training split is unknown, so the episodes we call held out may have been seen by these models during training.

Push world. The scene (MuJoCo, 64×64 frames from a fixed oblique camera) contains a walled table whose tint changes per episode, a position-controlled cylindrical pusher (a model step moves its set-point by up to 3 cm in each direction), a free block that interacts with the pusher through contact and receives a random horizontal force at every model step, and three kinematic distractor balls behind a divider that follow AR(1) processes with innovation 0.05, 0.3 and 1. The slowest is hidden by the divider in every frame, an error in the scene design that we found after training, so distractor retention is reported for the two visible balls only. The disturbance strength was chosen, on generated data only, so that the block's innovation given the actions (0.03–0.05) is comparable to that of Cube's contact-related variables; the prospectively specified window for it was widened once, before any model result, because an integrated contact-driven position cannot reach the originally registered range without jumping several centimetres per step. Estimated innovation of every factor is reported in Table 4. The data are 4 000 episodes of 50 model steps collected with a smooth random pushing policy that heads for the block half of the time; 400 episodes are held out. The model has a four-layer convolutional backbone (256 features), a batch-norm MLP projector to a 16-dimensional latent and an MLP predictor over three latents and action embeddings; it is trained for 40 000 steps with the LeWM loss and SIGReg weight 0.09 (batch 128, AdamW 10^{-3}, warm-up

and cosine decay), end to end, with a stop-gradient target, or with an inverse-dynamics term (weight 1). We evaluate the final checkpoint, as prospectively specified. Planning uses CEM (300 samples, 30 iterations, 30 elites) on the latent distance to a goal frame, replanning after every executed step in the simulator; every model faces the same 200 starts, goals and disturbance sequences. Pusher goals ask for the pusher within 2 cm of a point 5–10 cm away in 8 steps; block goals ask for the block within 3 cm of a point 5–10 cm away in 12 steps, which requires pushing it. Under the training disturbance the block drifts by several centimetres over that horizon, so block goals are evaluated in a quiet test world (the same scene and contact dynamics without the random force, also during the three context frames). The goal image of a block goal shows the block at its target and either the pusher at its starting position or the pusher where a straight push ends, 4.5 cm behind the target; the second aligns the pusher target with a straight push. Each condition was recorded before its corresponding evaluation; the second was added after inspecting results with the first (Appendix D). The goal images render the block at a fixed orientation, which adds an unreachable term for models that encode orientation. References on the same starts are random actions, zero actions, a scripted controller that reads the true state, walks around the block to the side opposite the goal and pushes it (it solves 83% of the quiet block goals), and, post hoc, a controller that ignores the block and drives the pusher to its position in the goal image.

Planning in the benchmark environments. Cube and Push-T follow the LeWorldModel protocol: CEM with 300 samples, 30 iterations and 30 elites over five blocks of five actions with receding horizon, goals taken from the same demonstration 75 steps (Cube) or 50 steps (Push-T) after the start, a budget of twice the goal offset, and success by the environment's own criterion (Cube: the cube within 4 cm of its goal position; Push-T: agent and block within 20 pixels and the block angle within $\pi/9$). Cube results use 300 fixed starts (100 historical starts and 200 registered fresh starts, the same for every model), Push-T 200 starts, and TwoRoom 100 starts with goals 25 steps ahead (success: the agent within 16 pixels of the target). Most Cube starts are windows of training episodes; the held-out control of §4 uses 100 starts from the held-out episodes of the paired models' split. Some evaluations simulate the starts in batches, which changes the planner's random draws but not the starts, goals or budgets; the default model was evaluated both ways, and the difference (-2 $[-9, +5]$ points on the first 100 starts, -0.3 $[-4.0, +3.3]$ on all 300) quantifies the observed effect of batching when interpreting comparisons that use different implementations.

Statistics. Seed variation is reported as ranges or standard errors over independently trained models or heads; single-seed results are marked. Planning differences between two models are paired by start and bootstrapped over starts (10 000 resamples, keeping all seeds of a start together), so the intervals condition on the trained models and measure start-set uncertainty, not training-seed variation; exact McNemar tests are given where one pair of models is compared. The correlation across suite settings (Appendix C.4) is a Spearman rank correlation over setting means, with one interval from resampling settings and a second, two-level interval from resampling the planning outcomes of each setting as well.

C Additional results

C.1 Appended signals

Table 3 summarises decodability of the appended signals. Two checks assess the protocol. First, the gripper opening, which is not appended, has similar decodability from the augmented and plain heads on the same encoder (0.94 and 0.92). Second, the empirical innovation of each synthetic stream, estimated from its own history, matches its design to 10^{-3}. The action-determined signal needs care: demonstrator actions are partly predictable from the frames (ridge R^2 0.42–0.45 from the current frame's features, 0.72 from three past frames), so that signal is not wholly new information for a model that sees the frames; the shuffled-action control, in which the predictor receives the actions of other windows, still keeps it. We therefore do not read its retention as evidence for action conditioning. One head trained with the weak regulariser weight also lost decodability of the constant signal (post hoc). The cause of this decline was not established. A post hoc nonlinear check on the six heads of LeWM's Cube encoder (a two-hidden-layer MLP probe, 256–256, trained on the same probe frames; recomputed ridge values match Table 3 to 0.003) agrees with the ridge probe: end to end, every AR signal with $q \geq 0.05$ has MLP $R^2 < 0$ in every head (-0.45 to -0.64), the

locked joint's jitter -0.39 to -0.45, while the constant, action-fixed, contact and cube-height channels stay at 0.95–0.99; with a stop-gradient target every channel is ≥ 0.99. Agreement between the probes supports reduced decodability beyond the linear probe class; it does not establish that the signals are unrecoverable by every possible decoder. On the encoder of our full-scale inverse-dynamics run, an end-to-end head keeps signals up to $q = 0.05$ (≥ 0.99) and removes those from $q = 0.1$ on (≤ 0.00), and a stop-gradient head keeps all (≥ 0.95; one head each): the removal pattern holds, at a slightly higher threshold than on the default encoder (the registered prediction, removal from $q = 0.05$, is not met).

Table 3: **Appended signals: retention** (R^2 from the latent, mean over heads; number of independently trained heads in parentheses; † a single head). The untrained head is the same measurement at initialisation.

head	q=0	q=0.02	q=0.05	q=0.1	q=0.2	q=0.35	q=0.5	q=0.75	q=1	action	locked jnt	contact	cube z
LeWM backbone, untrained head (3)	0.91	0.91	0.91	0.92	0.92	0.92	0.91	0.91	0.92	0.92	0.92	0.94	0.97
LeWM backbone, end to end (3)	0.99	0.66	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.98	0.01	0.87	0.99
LeWM backbone, stop-gradient (3)	0.99	0.98	0.98	0.98	0.98	0.98	0.98	0.96	0.95	0.97	0.94	0.96	0.99
LeWM backbone, e2e, λ=0.0117† (1)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.35	0.01	0.88	1.00
LeWM backbone, e2e, λ=0.45† (1)	0.99	0.99	0.98	0.99	0.97	0.52	0.01	0.00	0.00	0.98	0.02	0.90	0.99
LeWM backbone, e2e, λ=2† (1)	0.98	0.98	0.98	0.99	0.98	0.98	0.97	0.92	0.45	0.97	0.65	0.96	0.98
LeWM backbone, e2e, shuffled actions (3)	1.00	1.00	0.97	0.32	0.00	0.00	0.00	0.00	0.00	0.86	0.24	0.95	0.98
weak backbone, end to end† (1)	1.00	1.00	0.99	0.01	0.00	0.01	0.01	0.00	0.00	1.00	0.04	0.96	1.00
weak backbone, stop-gradient† (1)	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.98	0.97	0.98	0.94	0.98	0.99
DINOv2 backbone, end to end (3)	1.00	0.34	0.00	0.01	0.00	0.00	0.00	0.00	0.00	1.00	0.04	0.98	1.00
DINOv2 backbone, stop-gradient (3)	0.99	0.99	0.99	0.99	0.99	0.98	0.98	0.96	0.87	0.94	0.83	0.94	0.98

C.2 PUSH WORLD

Table 4: **Push world.** Top: estimated innovation fraction of each factor given the past and the actions, measured on the generated data. Middle: retention as MLP-probe R^2 at the end of training, from the 16-dimensional latent (the registered measurement) and from the encoder's 256-dimensional backbone; mean, range over models and number of models. Bottom: planning success on 200 shared starts per goal type — pusher goals in the quiet test world and under the training disturbance, block goals in the quiet world with a goal image that shows the pusher at its start or where a straight push ends — with references on the same starts. *Post hoc.

	pusher	block	tint	ball q=.3	ball q=1
innovation q (given actions)	0.006	0.041	0.000	0.307	1.002
end to end, latent	0.99 [0.99, 0.99] (3)	0.61 [0.59, 0.64] (3)	1.00 [1.00, 1.00] (3)	-0.07 [-0.07, -0.07] (3)	-0.05 [-0.06, -0.05] (3)
end to end, backbone*	0.99 [0.99, 1.00] (3)	0.91 [0.88, 0.93] (3)	1.00 [1.00, 1.00] (3)	-0.10 [-0.11, -0.09] (3)	-0.09 [-0.10, -0.09] (3)
detached target, latent	0.99 [0.99, 0.99] (3)	0.99 [0.99, 0.99] (3)	0.99 [0.99, 0.99] (3)	-0.16 [-0.18, -0.16] (3)	-0.13 [-0.14, -0.12] (3)
detached target, backbone*	0.99 [0.99, 0.99] (3)	0.99 [0.99, 0.99] (3)	1.00 [1.00, 1.00] (3)	0.59 [0.57, 0.63] (3)	0.06 [-0.10, 0.23] (3)
inverse dynamics, latent	1.00 [1.00, 1.00] (3)	0.85 [0.82, 0.86] (3)	1.00 [1.00, 1.00] (3)	-0.09 [-0.09, -0.09] (3)	-0.06 [-0.07, -0.05] (3)
inverse dynamics, backbone*	1.00 [1.00, 1.00] (3)	0.97 [0.97, 0.97] (3)	1.00 [1.00, 1.00] (3)	0.46 [0.28, 0.62] (3)	-0.06 [-0.07, -0.04] (3)
untrained encoder, backbone*	0.98 [0.97, 0.98] (2)	0.98 [0.97, 0.98] (2)	0.99 [0.99, 0.99] (2)	0.87 [0.85, 0.89] (2)	0.93 [0.92, 0.94] (2)

	pusher goals (quiet)	pusher goals (disturbed)	block goals (start image)	block goals (push image)
end to end	0.96 [0.95, 0.97] (3)	0.97 [0.96, 0.97] (3)	0.06 [0.05, 0.07] (3)	0.13 [0.10, 0.14] (3)
detached target	0.94 [0.93, 0.94] (3)	0.68 [0.56, 0.79] (3)	0.06 [0.06, 0.07] (3)	0.21 [0.20, 0.24] (3)
inverse dynamics	0.97 [0.96, 0.97] (3)	0.99 [0.98, 0.99] (3)	0.05 [0.04, 0.05] (3)	0.15 [0.12, 0.17] (3)
random actions	0.05	0.06	0.05	0.05
zero actions	0.02	0.03	0.04	0.04
true-state controller	0.99	—	0.83	0.83
block-blind controller*	—	—	0.04	0.13

Locating the reduction in block decodability. Probing the encoder's backbone as well as the latent (post hoc; all models re-probed with one code version on every fifth frame of the 400 held-out episodes) shows that end-to-end block decodability is higher in the backbone ($R^2 = 0.91$, range 0.88–0.93) than in the latent (0.61); pusher and tint decodability remain high in both. This resembles the backbone–latent gap of the weak-regulariser Cube models; an untrained encoder's 16-dimensional latent also has lower block decodability than its backbone (0.53 vs. 0.98), so the gap alone does not establish removal — the decline over training does (Fig. 3c). The untrained encoder's backbone supports decoding of every visible factor, including the distractor balls ($R^2 \geq 0.85$). After training neither ball is decodable from the latent; end to end, neither is decodable from the backbone.

A stop-gradient target and the inverse-dynamics term preserve some backbone decodability of the slower visible ball (0.59 and 0.46). The fast ball is not decodable by the MLP probe from any backbone (a ridge probe reaches $R^2 \approx 0.3$ on the stop-gradient and inverse-dynamics backbones). The slower visible ball was decodable from the stop-gradient latent early in training (MLP R^2 0.96 at 1 000–2 000 steps, 0.67 at 4 000, below zero from 8 000 on; one re-run) while the end-to-end latent had low early decodability (0.12 at 1 000 steps): content can be acquired and subsequently lose decodability without target gradients. The target-gradient mechanism of Prop. 2 does not explain this decline.

Why block goals are hard. Three post hoc diagnostics identify a contribution of the goal cost to planning difficulty; they do not exclude dynamics error, search limitations or goal infeasibility. (i) With goal images that show the pusher at its start, the encoded true outcome of a successful scripted push is farther from the goal image than the start frame in 90–100% of cases for every model, and each model predicts a higher cost for the scripted push than for doing nothing: the pusher term dominates the distance. With goal images that show the pusher where a push ends, the ordering is right in 70–89% of cases. (ii) A 5 cm block displacement produces a pairwise squared distance between the original and perturbed scene encodings that is 0.3–1% of the distance produced by an equal pusher displacement in end-to-end and inverse-dynamics models, compared with 34–48% for a stop-gradient target (Fig. 3f). These pairwise distances differ from the fixed-goal cost maps in Fig. 3e; the per-episode tint dominates every latent but is constant within an episode. (iii) A controller that ignores the block and only drives the pusher to where the goal image shows it solves 0.13 of the push-image goals — it pushes the block along whenever it happens to start behind it — and only the stop-gradient comparison has an interval excluding zero (paired over starts, pooled over three models: stop-gradient $+0.08$ $[+0.04, +0.13]$, inverse dynamics $+0.02$ $[-0.03, +0.06]$, end to end $+0.00$ $[-0.05, +0.05]$). The goal images also render the block at a fixed orientation, which adds an unreachable term for models that encode orientation. Under the training disturbance, the stop-gradient models reach pusher goals less often than the others (0.68 vs. 0.97–0.99) and act more forcefully (mean $|a|$ 0.69 vs. 0.42), consistent with the drifting block entering their cost.

C.3 Full-scale retraining, the state-aligned model and TwoRoom

Variance associated with state variables. Probe accuracy measures decodability, whereas Euclidean distances also depend on how the representation varies with a factor. Table 5 reports the latent variance explained by arm and contact-related variables, using least squares on standardised factors and their squares (post hoc). These groups overlap and need not sum to one; they do not constitute an exhaustive decomposition of all state information. In the default and weak LeWM models, the contact-related variables explain about 2% and the arm 7–19%. The state-aligned model associates a larger share with these variables (arm 56%, contact 15%) and has a lower participation ratio (7.6, compared with about 100 for LeWM). These measurements describe the learned representation, but do not identify a causal contribution to planning or establish that the induced distance matches the task objective.

Table 5: **Latent variance explained by state-variable groups on Cube** (post hoc). Fractions are variance-weighted R^2 values from regression of standardised state variables and their squares onto the latent. The groups overlap. Ranges are over trained models.

model	arm	contact-related variables
LeWM, default	0.07	0.02
LeWM, weak regulariser	0.18–0.19	0.02
LeWM + inverse dynamics	0.12	0.03
State alignment + inverse dynamics (Liu et al., 2026)	0.56	0.15

Training dynamics. Figure 5 follows the full-scale runs epoch by epoch, with the same probe evaluation applied to each epoch's checkpoint. LeWM acquires the contact-related state over the first epochs and keeps a small backbone–latent probe-score gap throughout. With inverse dynamics, the state is highly decodable from both representations after one epoch; after ten epochs, contact-related probe R^2 is 0.975 in both (LeWM: 0.92 in the latent, 0.94 in the backbone), at a normalised one-step error of 4.8×10^{-3} (LeWM: $(4.7\text{–}5.2) \times 10^{-3}$). Its latent has a modestly larger variance share

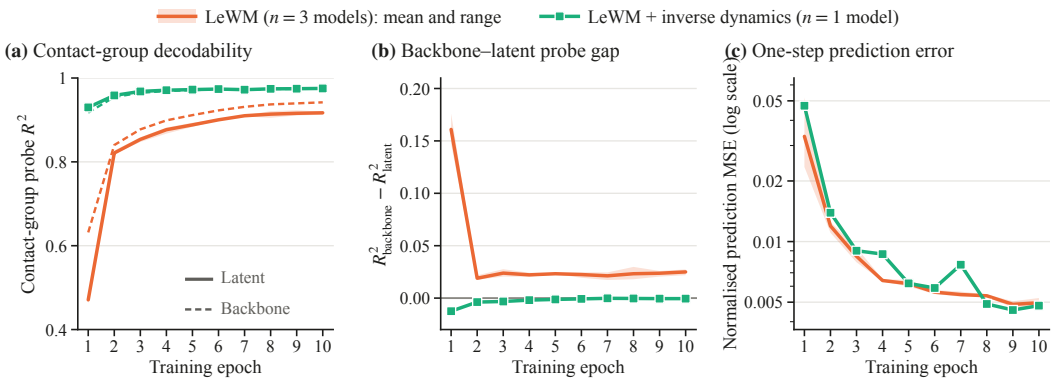

Figure 5: **Contact-related state during full-scale training on Cube.** (a) Probe R^2 of the contact-related factors from the latent (solid) and the encoder's backbone (dashed); (b) backbone–latent probe-score gap; (c) normalised one-step prediction error. LeWM (three runs: mean and range) acquires the contact-related variables over several epochs and maintains a small backbone–latent gap. With inverse dynamics (one run), both representations support high decodability from the first epoch; after ten epochs, prediction error is within the default-run range.

associated with the measured state variables than LeWM's (Table 5). The planning improvement is based on one trained model. The lower confidence bound is $+3.0$ points, so an improvement of at least five points is not established at the reported confidence level.

External model on Push-T. The state-aligned model reaches 55.0% planning success versus 40.5% for the default model on 200 shared starts, a paired difference of $+14.5$ points $[+7, +22]$. Block-angle decodability is 0.84 versus 0.81. These evaluations use our planning protocol, and the external model's training split is unknown.

TwoRoom. In TwoRoom the only state is the agent's position, which has almost no innovation given the actions. Every model supports accurate position decoding (Table 6), consistent with the low-innovation setting of the theory, and the three retrained heads plan within five points of each other (85–90%); the state-aligned model of Liu et al. (2026) reaches 100% and the public LeWM model 81% ($+19$ $[+12, +27]$). Their different planning outcomes therefore cannot be explained by position decodability alone.

Table 6: **TwoRoom** (100 shared starts, goals 25 steps ahead). Decodability of the agent's position from the latent and the backbone, one-step prediction error, and planning success. The heads are retrained on the fixed encoder of the public LeWM model (one head each).

model	agent position R^2 (latent)	(backbone)	one-step error (10^{-3})	planning success (%)
LeWM (public model of its authors)	0.992	0.994	65.8	81
state alignment + inv. dyn. (Liu et al., 2026)	1.000	1.000	9.9	100
retrained head, end to end	0.996	0.994	12.3	85
retrained head, detached target	0.994	0.994	27.2	88
retrained head, inverse dynamics	0.998	0.994	13.2	90

C.4 Model suite

The complete model collection contains 22 Cube and 3 Push-T LeWM models trained with the procedure of Appendix B, plus the authors' public models. On Cube (10 000 expert episodes of the single-cube task): the default weight $\lambda = 0.09$ and a weak weight $\lambda = 0.0117$ (three seeds each), five further weights between 0.03 and 0.32 (one seed each) and $\lambda = 0.45$ (two seeds), predictor depth one and three, dropout, a wider predictor MLP, a predictor left at initialisation, and a mixture-density prediction head (three seeds); plus the authors' public model. On Push-T: the default and weak weights, the mixture head, and the public model; these runs used LeWM's window-level split, so the probes' test episodes were seen during world-model training (the probes' own split is by

episode). Every measurement uses the exported weights supplied to the planner. Probes are ridge regressions on standardised features, with the regularisation chosen on an inner split of the training frames, evaluated on episodes held out from probe fitting; features are the 192-dimensional latent or the encoder's CLS token.

Per-factor decodability under weak regularisation. From $\lambda = 0.09$ to 0.0117 (three seeds each, epoch-10 latents), the gripper opening falls from 0.91–0.93 to 0.43–0.59, the contact flag from 0.83–0.85 to 0.37–0.55 and the cube height from 0.99 to 0.76–0.88, while the arm joints ($0.97 \rightarrow 0.99$) and the cube's planar position ($0.99 \rightarrow 0.98$) are kept; the backbone keeps 0.72–0.81, 0.65–0.75 and 0.93–0.97 of the three factors. A two-hidden-layer MLP probe corroborates the lower contact-related decodability (e.g. contact group 0.36 vs. 0.52 from one weak latent, 0.82 vs. 0.77 from its backbone). Across-model differences do not establish monotonic forgetting during a run: the weak-weight models show a 0.09–0.12 decline from epoch 1 to epoch 5, followed by recovery, with maximal retention at epoch 10. The prespecified final-versus-peak decline was therefore not observed (P2c, Appendix D).

Content and planning across the suite. Across the 13 Cube settings of the suite (eight regulariser weights, two predictor depths, dropout, a wider predictor MLP and the mixture-density head; setting means over seeds; planning on 300 fixed starts), contact-related decodability correlates with planning success (Spearman $\rho = 0.81$; the interval is $[0.41, 0.97]$ when settings are resampled and $[0.10, 0.88]$ when planning outcomes are resampled as well), whereas 10-step rollout error has a weak observed correlation (-0.10). The suite is small and heterogeneous, the settings were not chosen for this test, and in the head experiments the participation ratio tracks planning as well ($\rho = 0.78$ across 20 heads, against 0.50 for contact decodability); we therefore interpret these correlations as descriptive associations.

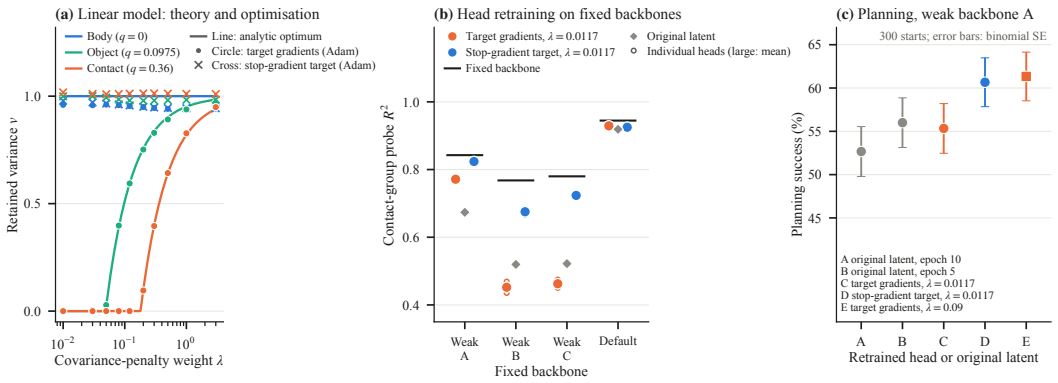

Figure 6: **Linear theory and head retraining.** (a) Linear JEPA: retained variance of three factors predicted by equation 3 (lines) and obtained by numerical optimisation with target gradients (circles) or a stop-gradient target (crosses). (b) Contact-related decodability after retraining only the head on a fixed backbone, for the three weak-regulariser backbones (A–C) and a default one; small markers are individual heads (two per objective on the weak backbones, one on the default) and large markers their mean; black line: the fixed backbone; diamond: the original model's latent. (c) Planning success on 300 fixed starts for weak backbone A: the original model at its final and best early epoch, and three retrained heads. Error bars: binomial standard error over the 300 starts; one model or head per point.

What the mechanism does not explain in the suite. Contact-related decodability does not keep rising above the default weight; retention and planning do not move together within runs; the dataset innovation ordering fails for cube height and position; and the sensitivity of planning to distribution shift, the surprise signal and longer head training did not behave as registered (Appendix D). The benchmark variables are statistically dependent: contact also predicts cube motion, so gradients through the context can favour its retention. The appended-signal experiments exclude this dependence by construction.

C.5 Experiments not completed

Two planned full-model experiments were not completed within the available compute budget: LeWM on Cube with the target detached (registered as P8e) and LeWM on Push-T with the inverse-dynamics term (planned, not registered). Each would use the training procedure and evaluation of the completed full-scale run (about twelve hours on one 24-GB GPU); the public code contains both configurations, and the paper makes no claim that depends on them.

D Prospectively recorded predictions and outcomes

The tables below distinguish predictions recorded before the corresponding evaluation from analyses added after observing results. Predictions, thresholds and statistics were committed to version control; the supplementary material provides the record and timestamps. The record also identifies prior evidence available when a prediction was specified, including preliminary results on some models and published results from other groups. Thus, these entries form a sequence of prospective tests, with disclosed prior evidence and protocol amendments, rather than a single protocol fixed before the entire study. The tables cover the analyses reported here; the accompanying record also retains predictions and outcomes for analyses omitted from the manuscript.

Protocol amendments. Three changes were made to the pushing experiments before the evaluations to which they applied. First, the block's innovation range was widened after generating the data but before evaluating any trained model: achieving the original range would have required an integrated, contact-driven position to jump several centimetres per step. Second, quiet-world block-goal tests were added after the initial disturbed-world evaluation showed very low success. Third, goal images placing the pusher at the end of a straight push were added after the first quiet-world tests showed that goal images fixing it at its start penalised pushing. Both quiet-world goal-image settings are reported, with this sequence made explicit; their prospective recording does not make the amended design independent of earlier observations.

Exploratory analyses and incomplete tests. Analyses added after inspecting results include the nonlinear probes of appended signals, the push-world backbone probes and distance-sensitivity measurements, the block-blind reference controller, the variance explained by state variables, and additional planning diagnostics and zero-action references. These are identified as exploratory in the main text or figure captions. The full-model stop-gradient experiment (P8e) was recorded but not performed; its row is retained. The planned full-model inverse-dynamics experiment on Push-T was neither recorded prospectively nor run (Appendix C.5).

Reading the tables. Identifiers match the accompanying record. "Not met (not falsified)" indicates that a prespecified success criterion was not met, while a separately specified falsification criterion was also not met. This distinction does not provide positive evidence for the prediction. Numerical success criteria concern the stated statistic; meeting a point-estimate threshold does not establish the same lower bound on the population effect. For example, P8c's observed improvement is 7.7 percentage points, but its interval $[3.0, 12.3]$ does not establish an improvement of at least five points. The PB1 retention-gap threshold is met by two of the three weak backbones. Seed counts and the scope of uncertainty are stated in Appendix B. In the tables, "e2e" denotes training with target gradients and "sg" denotes a stop-gradient target.

Table 7: Prospectively recorded predictions for the model suite and head retraining.

id	prediction	observed	outcome
P1	dataset innovation: body $<$ planar object $<$ lift, gripper, contact (Cube); agent $<$ block $<$ angle (Push-T)	Cube: contact 0.067, gripper 0.039 $>$ cube x, y; but arm joint 0.020 $>$ lift 0.015 (falsifier fires), $>$ cube x 0.019 and y 0.006, and cube $x >$ lift; Push-T $0 < 0.01 < 0.18$	Cube no; Push-T yes
P2a	$\rho(q, \Delta R^2) \leq -0.5$ (Cube), negative (Push-T)	Cube $\rho = +0.01$ [-0.63,+0.57]; Push-T $\rho = -0.93$	Cube no; Push-T yes
P2b	$\rho(\lambda, \text{contact } R^2) \geq 0.6$; arm $R^2 \geq 0.95$ everywhere	$\rho = +0.33$ [-0.67,+0.90]; min arm R^2 0.87	no (inverted U)
P2c	weak-λ runs lose ≥ 0.1 contact R^2 from their peak by epoch 10	dip then recovery (epoch 1$\rightarrow$5 drop 0.09–0.12; epoch 10 is the maximum)	no
P2d	$\rho(\lambda, \text{one-step NMSE}) \geq 0.8$	$\rho = +0.98$ [+0.73,+1.00]	yes
P2e	$\rho(\text{contact } R^2, \text{success}) \geq 0.5$ and $> \lvert\rho(\text{rollout error, success})\rvert$	$\rho = +0.81$ [+0.41,+0.97] vs $\rho = -0.10$ [-0.87,+0.52]	yes
PB1	backbone$-$latent contact gap ≥ 0.2 (weak), ≤ 0.1 (default)	weak 0.17/0.25/0.26; default 0.03/0.02/0.03	weak 2/3; default yes
P3a	head retraining, weak backbone A: end-to-end λ=0.0117 keeps ≥ 0.15 less contact R^2 than end-to-end λ=0.09	0.77 vs 0.82 (difference 0.05)	no
P3b	stop-gradient head within 0.05 of the backbone; higher one-step error than end-to-end	gap 0.02 (backbone A); replicates 0.10, 0.06; NMSE 3–4$\times$ higher	backbone A only
P3c	a content-keeping head $\geq +5$ pp over the original weak model; default-backbone heads within ± 5 pp	+14/+7/+11 pp on the three weak backbones; default heads -4 to $+5$	yes as registered (n=100); confounded with early stopping (P3f)
P3f	stop-grad head $\geq$ own epoch-5 model $+5$ pp (300 starts)	$+4.7$ [$+1.0, +8.3$], $p = 0.02$	no (borderline)
P3g	stop-grad $-$ end-to-end $\geq +3$ pp on 200 fresh starts, pooled over three backbones	$+5.3$ [$+3.0, +7.8$], 47 vs 15; backbone A known at registration; unseen backbones only $+4.75$ [$+2.0, +7.8$]	yes
P3h	second head seed: stop-grad keeps more contact than end-to-end on $\geq$ 2 of 3 backbones; pooled planning difference $\geq +3$ pp	3 of 3 backbones (0.83 vs 0.77, 0.68 vs 0.47, 0.73 vs 0.45); $+6.3$ [$+2.0, +11.0$], 29 vs 10	yes
P3i	Push-T (goal offset 50, 200 starts): stop-grad $-$ end-to-end head $\geq +3$ pp	-2.0 [$-8.5, +4.5$], 21 vs 25; weak original 49.0% vs default 40.5%	no

Table 8: Prospectively recorded predictions for appended signals and the generic encoder. †One trained head per condition.

id	prediction	observed	outcome
P5a	appended AR signals, e2e: $\rho(q, R^2) \leq -0.7$ per seed and $R^2(q{=}1) \leq R^2(q{=}0) - 0.3$	ρ -0.97/-0.63/-0.82; gap 1.00/1.00/1.00	not met (not falsified)
P5b	stop-gradient keeps q=1 signal: sg$-$e2e ≥ 0.2 per seed	0.96/0.96/0.95	yes
P5c	action-determined signal kept: R^2(action) $> R^2(q{=}0.5)$ per seed	differences 0.99/0.98/0.99	yes (confounded; see P5j)
P5d	locked-joint velocity (q=0.73): sg$-$e2e ≥ 0.1 per seed	0.93/0.93/0.92	yes
P5e†	λ=0.0117 removes mid-q signals more than 0.09	mean R^2 0.00 vs 0.00; at 0.0117 even q=0 is removed (0.00) and the action signal falls to 0.35	yes by the rule; both ≈ 0
P5f†	weak backbone: P5a ρ and P5b gap hold	ρ -0.80; gap 0.97	yes
P5g	half-life of removal decreasing in q: $\rho \leq -0.8$ per seed	ρ -0.97; ρ -0.93 (a never-halved signal scored as longest, not stated at registration; without it -0.89)	yes
P5h	gripper-velocity sensor channel: sg$-$e2e ≥ 0.1 per seed	e2e 0.01/0.01/0.00, sg 0.98/0.98/0.98 (3 seeds)	yes
P5i†	strong regularisation (λ=0.45, 2.0): q=1 and $q \geq .35$ signals ≤ 0.2, q=0 ≥ 0.9	λ=0.45: q=1 -0.00, $q \geq .35$ 0.13, q=0 0.99; λ=2: q=1 0.45, $q \geq .35$ 0.83, q=0 0.98	not met (not falsified)
P5j	predictor given other windows' actions: action-determined signal removed (≤ 0.3), q=0 kept	action-determined 0.86/0.88/0.84; q=0 1.00/1.00/1.00 (3 seeds)	no
P7a	DINOv2 backbone, e2e: AR signals with $q \geq 0.05$ removed, q=0 and action kept	q=0 1.00, max($q \geq .05$) 0.00, action 1.00; q=0 1.00, max($q \geq .05$) 0.01, action 1.00; q=0 1.00, max($q \geq .05$) 0.01, action 1.00 (3 seeds)	yes
P7b	DINOv2 backbone, sg: every AR signal ≥ 0.8	min 0.88; min 0.87; min 0.86 (3 seeds)	yes
P7c†	DINOv2 heads: sg contact $\geq$ e2e -0.02	sg 0.74 vs e2e 0.78 (backbone 0.75)	not met (not falsified)

Table 9: Prospectively recorded predictions for full-model retraining, the state-supervised model of Liu et al. (2026), TwoRoom, the push world and inverse-dynamics heads. †A single model or head. *The block-blind reference was added post hoc.

id	prediction	observed	outcome
P8a	full run with inverse dynamics: contact-driven $R^2 \geq 0.94$, backbone gap ≤ 0.03	0.98 (default 0.92); gap -0.00	yes
P8b	one-step error higher than the default (registered: three-seed mean)	NMSE 0.0048 vs 0.0050 (same seed 0.0047)	no
P8c	planning, 300 Cube starts: $\geq +5$ points over the default	+7.7 [+3.0, +12.3] (69.3% vs 61.7%)	yes
P8d	appended signals on the inverse-dynamics backbone: same removal pattern	e2e max($q \geq .05$) 0.99, detached min 0.95	not met (not falsified)
P8e	full run with detached target: content $\geq$ default; one-step error higher; no planning prediction	not run (no second GPU before the deadline)	—
P9a	state-aligned model of Liu et al. (2026): contact ≥ 0.94, gap ≤ 0.03	0.98; gap -0.01	yes
P9b	planning $\geq$ default (63% on 100 starts)	83% vs 63% (100 starts; +20 [+12, +28]); 300 starts 81.3% (+19.7 [+15.0, +24.3])	yes
P9c	state-aligned model on Push-T: block angle ≥ 0.85; planning $\geq$ default	angle 0.84 (default 0.81); planning 55.0% vs 40.5% (+14.5 [+7.0, +22.0])	not met (not falsified)
P9d	appended signals on the state-aligned model's encoder: same pattern	e2e max($q \geq .05$) 0.99; detached min 0.96	not met (not falsified)
P10a	TwoRoom encoder: end to end removes $q \geq 0.05$ (≤ 0.2), detached keeps all (≥ 0.8)	e2e max($q \geq .05$) 1.00; detached min 0.99	no
P10b	TwoRoom heads: detached and inverse dynamics within -3 points of end to end; agent $R^2 \geq 0.95$	e2e 85%, detached 88%, inv. dyn. 90%	yes
P11a	push world, end to end: pusher, tint ≥ 0.9; block ≤ 0.3; q=1 ball ≤ 0.1	block 0.60/0.64/0.59	not met (not falsified)
P11b	push world, detached target: slow ball ≥ 0.8; block ≤ 0.3	block 0.99/0.99/0.99; slow ball (occluded) -0.19/-0.22/-0.18	no
P11c	push world, inverse dynamics: block ≥ 0.8; block goals $\geq$ end to end $+0.2$; pusher goals ≥ 0.8 for all	block 0.82/0.86/0.85; block-goal diff +0.01 [-0.01, +0.02]; pusher goals min 0.56	not met (not falsified)
P11d	quiet world, block goals: success rises with block retention ($\rho \geq 0.5$); models with block $R^2 \geq 0.8$ beat random by ≥ 0.1	$\rho = 0.09$ [-0.79, 0.84] (9 models); random 0.05, scripted 0.83	not met (not falsified)
P11c′	inverse dynamics vs end to end, block goals in the quiet world: $\geq +0.2$ (falsified if ≤ 0)	-0.01 [-0.03, +0.00] (3 paired seeds)	no
P11e	push-end goal frames, quiet world: success rises with block retention ($\rho \geq 0.5$); detached $>$ end to end; block $R^2 \geq 0.8$ beats random by ≥ 0.1	$\rho = 0.71$ [-0.17, 0.98] (9 models); detached 0.21, end to end 0.13; random 0.05, block-blind* 0.13, scripted 0.83	not met (not falsified)
P12a	inverse-dynamics heads, weak backbones: contact higher than end-to-end heads on $\geq 2/3$	0.89 vs 0.77/0.84 vs 0.44/0.85 vs 0.47	yes
P12b	same: planning vs end-to-end heads, pooled $\geq +3$ points	+7.3 [+2.0, +13.0] (batched vs. unbatched planner; vs. detached heads +0.0 [-4.0, +4.0])	yes
P12c†	Push-T weak backbone: inverse-dynamics head keeps more block angle; planning (directional)	0.72 vs 0.72; planning 47.5% vs 46.5% (+1.0 [-4.5, +7.0])	not met (not falsified)

E EXPERIMENTS WITH A LINEAR JEPA

Setup. Eight independent unit-variance factors evolve as $x_i' = \rho_i x_i + g_i a_i + \sqrt{q_i}\,\epsilon_i$ with independent standard-normal commands a_i and noise ϵ_i: two body factors ($\rho = 0.6$, $g = 0.8$, $q = 0$), a planar object factor ($\rho = 0.95$, $q = 0.0975$), a lift factor ($\rho = 0.9$, $q = 0.19$), a contact factor ($\rho = 0.8$, $q = 0.36$), two slow distractors ($\rho = 0.97$, $q = 0.059$) and pure noise ($q = 1$). The encoder sees a fixed random rotation of the factors (80 000 steps). A linear encoder $W \in \mathbb{R}^{8\times 8}$ (initialised at the identity), a linear predictor and a linear action map are trained jointly with Adam (3×10^{-3}), using the one-step prediction objective and λ times either the covariance penalty $\|\mathrm{Cov}\, z - I\|_F^2/D$ (12 000 steps, batch 1024) or LeWM's SIGReg module (15 000 steps, batch 128, as in LeWM training), end to end or with a detached target. Retained variance is the latent variance carried by each factor.

Results. With the covariance penalty, optimisation reproduces (Fig. 6a) equation 3 factor by factor and the normalised prediction error to within 0.002 for $\lambda \leq 1$ (e.g. $\lambda = 0.3$: 0.080 vs. 0.081; $\lambda = 0.12$: 0.043 vs. 0.043). Among the tested weights, factors are lost in the order noise ($\lambda \leq 0.5$), contact (≤ 0.12), lift, planar object and distractors; the body is never lost. SIGReg produces the same ordering (at $\lambda = 0.1$, noise retains 0.18 and contact 0.73 while the planar object retains 1.08), but retained factors can exceed unit variance. This is consistent with the regulariser constraining random one-dimensional projections, allowing variance lost in one direction to be partly compensated in others. With a detached target, every factor retains unit variance at every tested λ under both penalties, and the prediction error remains at the value when all factors are retained, ≈ 0.22.

F EXTENDED RELATED WORK

This appendix expands the positioning summarised in §5.

Predictability and feature selection. Temporal regularity is central to slow feature analysis and its nonlinear extensions (Wiskott & Sejnowski, 2002; Sprekeler et al., 2014), predictable feature analysis (Richthofer & Wiskott, 2015), dynamical components analysis (Clark et al., 2019), and the past–future information bottleneck (Creutzig et al., 2009). Related analyses establish biases towards slow or predictive features in JEPAs (Sobal et al., 2022; Littwin et al., 2024). In self-predictive reinforcement learning, theoretical studies connect representation learning to spectral properties of the transition operator and examine the effects of optimisation dynamics, stop-gradients and distractors (Tang et al., 2023; Ni et al., 2024; Voelcker et al., 2024; Khetarpal et al., 2025). Proposition 4 specialises this established selection principle to independent linear-Gaussian factors with actions. Our emphasis is the contribution of gradients through the target to the suppression of already represented signals, its dependence on innovation, and its measurement during training.

Distributional constraints and identifiability. LeJEPA combines predictive learning with Gaussian regularisation (Balestriero & LeCun, 2025), and LeWorldModel applies this formulation to action-conditioned dynamics from pixels (Maes et al., 2026). Klindt et al. (2026) establish identifiability under assumptions on the latent distribution, transition dynamics and representation dimension. Their exact Gaussian result assumes matched state and embedding dimensions and isotropic stationary transitions; dimension matching alone is insufficient. Our linear analysis examines which factors are selected when the representation is restricted, and how a soft isotropy penalty trades retained variance against prediction error. Proposition 3 addresses a complementary limitation: a regulariser depending only on the marginal latent distribution cannot distinguish encoders that induce the same distribution. Its local statement about reintroducing an omitted signal requires independence and a smooth additive perturbation; it is not a general impossibility result for information preservation by regularisation.

Acquisition, suppression and auxiliary objectives. Pendharkar (2026) vary predictability independently of controllability and relevance, finding that the evaluated reward-free objectives fail to retain an exogenous task-relevant feature. Their endpoint measurements establish omission; our experiments additionally track signals represented before training and a contact-driven object acquired and subsequently suppressed. Multi-step inverse models seek representations of endogenous state while excluding exogenous distractors (Efroni et al., 2022; Lamb et al., 2023; Islam et al.,

2023). These objectives do not imply that all unpredictable information is irrelevant to control. Stop-gradient and moving-average targets are established in self-supervised and model-based learning (Grill et al., 2020; Schwarzer et al., 2021; Hansen et al., 2024). We use stop-gradient targets to isolate the target contribution, without claiming that they guarantee acquisition or preservation under arbitrary training dynamics. Recent JEPA variants add inverse dynamics (Ivashkov et al., 2026; Zhang et al., 2026), state supervision (Liu et al., 2026), proprioceptive grounding (Yan et al., 2026), or subspace regularisation (Zhao et al., 2026); our interventions examine how the base objective's behaviour changes with stop-gradient targets or inverse dynamics.

Representation geometry and planning. The mismatch between prediction accuracy and control performance motivates value-aware modelling and related objectives (Lambert et al., 2020; Wei et al., 2024; Grimm et al., 2020; Farahmand et al., 2017); bisimulation and latent-MDP formulations relate representations to behavioural relevance (Gelada et al., 2019; Zhang et al., 2021). For latent world models, recent studies examine prediction–planning discrepancies (You et al., 2026; Singh, 2026), rollout consistency (An et al., 2026), and the distinction between decodability and useful distance geometry (Hu et al., 2026; Li et al., 2026; Wang et al., 2026). We identify one source of this distinction: in the aligned linear solution, suppression reduces a factor's variance and its weight in the Euclidean cost. For nonlinear encoders, we measure cost sensitivity directly. This mechanism complements explanations based on dynamics error, reachability and optimisation. Spectrum-based statistics such as RankMe (Garrido et al., 2023) describe the representation's distribution but cannot identify which state variables it contains (Prop. 3(a)); Appendix C.4 evaluates their empirical relationship with planning.