# WHEN PREDICTION CHOOSES THE REPRESENTATION: LEARNING DYNAMICS AND PLANNING IN WORLD MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

World models are trained to predict but deployed to make decisions. What happens between these goals when an encoder and latent dynamics model learn together? Predictive training creates a competition: features that are already easy to represent provide stronger learning signals to their predictors, while harder decision-relevant features can lose representation and become still harder to learn. A model can therefore predict accurately and avoid collapse while presenting the wrong world to a planner. Constraints on marginal latent statistics cannot rule out this failure, because the same healthy representation can encode controllable structure or predictable nuisance. We formalize this feedback as a coupled selection process and identify a recoverability boundary: once the decision mode falls too far behind, an intervention of bounded strength can no longer reverse its decline. Before that boundary, the analysis yields the minimal directional protection needed to preserve a chosen share of decision information and keep its predictor learning. Controlled visual planning experiments reproduce the resulting separation between prediction and planning, the transition caused by stronger nuisance, and the importance of intervening in the right direction. The practical message is that anti-collapse should not merely keep the latent active; it should protect the information that actions and planning require before predictive training starves it of learning signal.

## 1 INTRODUCTION

World models learn to predict futures so that agents can choose actions (Ha & Schmidhuber, 2018; Hafner et al., 2020). Yet accurate prediction of a learned latent state does not ensure that the state retains what decisions require. A predictable background can satisfy the training objective while the controllable state disappears. Anti-collapse constraints keep representations informative, but do not determine which information survives (Bardes et al., 2022; Maes et al., 2026).

The difficulty begins during joint training. When a decision feature receives little representation, its predictor learns slowly. The remaining predictor error then makes that feature expensive to retain, encouraging the encoder to suppress it further. An easy nuisance benefits from the opposite feedback. Prediction can improve because training selects an easier part of the world, even as the information needed for planning becomes harder to recover.

We ask when a training constraint can interrupt this feedback. In a coupled encoder–predictor model, recoverability depends jointly on the surviving decision share, the predictor's remaining error, and the available intervention strength. When the nuisance predictor is already fitted, we derive an exact boundary separating states that a bounded directional intervention can recover from states that no intervention within the same bound can recover. The boundary makes timing a property of the learning state: the same intervention can succeed before suppression advances and fail afterward.

The positive result is a principle of targeted protection. Keeping the decision share $p$ above a chosen threshold $\tau$ guarantees that its predictor continues to learn. At the threshold, the least nonnegative score correction is

$$v_{\min} = (y - z - a)_+ \qquad \text{when } p = \tau, \tag{1}$$



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.

where $y, z$ are the decision and nuisance predictors' squared mismatches, $a$ is the decision mode's selection advantage, and $(x)_+ = \max\{x, 0\}$. The rule compensates precisely for the current disadvantage and is zero away from the threatened boundary. It therefore specifies the direction and minimum instantaneous strength of useful protection without prescribing a marginal latent distribution.

This perspective connects a training mechanism to a practical design choice. Constraints should respond to threatened decision information in the action-relevant direction. A separate geometry-to-planning bound identifies how the retained information must be exposed to the planner. Controlled pixel-world experiments with the LeWorldModel architecture examine information loss during training and compare marginal regularization with action-conditioned constraints (Maes et al., 2026).

**Contributions.**

- **A quantitative recovery boundary.** We derive coupled representation–predictor dynamics and an exact separatrix for bounded score intervention in the mature-nuisance regime, identifying when decision information remains recoverable.
- **Minimal protection with a learning guarantee.** We construct the least nonnegative boundary correction that preserves a specified decision share and prove an explicit predictor-error decay rate.
- **A constraint-design principle tied to planning.** We show that useful protection must act on threatened decision directions, connect the retained representation to planning error, and test these predictions with action-conditioned constraints.

## 2 Related Work

**World models and latent planning.** World models learn a compact state in which agents can predict and compare futures (Ha & Schmidhuber, 2018; Hafner et al., 2019). Dreamer learns behavior through imagined trajectories, and its later formulation scales across diverse control domains (Hafner et al., 2020; 2025). TD-MPC2 performs trajectory optimization with an implicit world model (Hansen et al., 2024). LeWorldModel jointly learns an encoder and action-conditioned predictor from pixels using next-embedding prediction and Gaussian regularization (Maes et al., 2026). This formulation makes the interaction between representation allocation and predictor learning central to training.

**Predictive representations and non-collapse.** BYOL and SimSiam use encoder–predictor asymmetry to learn without negative pairs (Grill et al., 2020; Chen & He, 2021). Barlow Twins reduces redundancy through cross-correlations, while VICReg controls variance and covariance (Zbontar et al., 2021; Bardes et al., 2022). I-JEPA predicts target-block embeddings from context (Assran et al., 2023). Analyses of predictive embeddings identify preferences for slow or high-influence features (Sobal et al., 2022; Littwin et al., 2024). In LeWM, temporally centered SIGReg separates persistent and residual variance to improve robot policy learning (Liu et al., 2026). Physical-parameter interventions further show that observation access and prediction targets determine which quantities a latent acquires (Tan et al., 2026).

**Feature learning and training transients.** Deep linear models provide explicit accounts of mode acquisition during learning (Saxe et al., 2014). Tian et al. analyze coupled encoder–predictor dynamics, collapse basins, and the curriculum induced by a moving target encoder (Tian et al., 2021). Gradient starvation describes how learning some predictive features can suppress others, and motivates decoupling their learning dynamics (Pezeshki et al., 2021). Regularization can be decisive during an early transient while having little effect near convergence (Golatkar et al., 2019); recent self-supervised studies connect learning phases to transferability and critical periods (Lee et al., 2026). Our analysis studies the resulting decision-information feedback through bounded recovery and guaranteed predictor learning under directional protection.

## 3 PROBLEM SETUP

We study a controlled process with history $h_t = (o_{\le t}, u_{<t})$, action $u_t$, and planning horizon $H$. When a state description is available, we write $x_t = (q_t, n_t)$, where $q_t$ affects decisions and $n_t$ is nuisance variation. A joint-embedding world model learns an encoder $E_\theta$ and latent predictor $F_\phi$ from

$$\mathcal{L}(\theta, \phi) = \mathbb{E}\Big[\|E_\theta(h_{t+1}) - F_\phi(E_\theta(h_t), u_t)\|^2\Big] + \lambda R(E_\theta). \tag{2}$$

We call $R$ *marginal-only* if it depends on $E_\theta$ only through the latent distribution, as do standard variance, covariance, and distribution-matching constraints.

Decision relevance is defined by action-conditioned futures. For an action sequence $\mathbf{u}$, let

$$K_H^{\mathbf{u}}(\cdot \mid h) = \mathbb{P}(Y_H \in \cdot \mid h, \mathrm{do}(\mathbf{u})). \tag{3}$$

Histories $h, h'$ are equivalent when these kernels agree for every $\mathbf{u}$. Their equivalence classes form the decision quotient $Q$, with quotient map $\pi : \mathcal{H} \to Q$. A planning representation $P$ is *decision-sufficient* if $\pi(h) = g(P(h))$ for some $g$. We allow $P = E_\theta$ or a head $P_\psi = H_\psi \circ E_\theta$. Given a task distance $d_Q$, $P$ is $(\alpha, \beta)$-faithful when

$$\alpha d_Q(\pi(h), \pi(h')) \le \|P(h) - P(h')\| \le \beta d_Q(\pi(h), \pi(h')), \qquad \kappa = \beta/\alpha.$$

For a start history $h$, let $q_H(\mathbf{u}; h)$ be the true quotient endpoint and $q_g$ the goal. We evaluate

$$J_Q(\mathbf{u}) = d_Q(q_H(\mathbf{u}; h), q_g), \qquad \widehat{J}(\mathbf{u}) = \|\widehat{y}_H(\mathbf{u}) - P_\psi(q_g)\|, \tag{4}$$

where $\widehat{y}_H = H_\psi(F_\phi^{(H)}(E_\theta(h), \mathbf{u}))$. We denote uniform rollout error by $\varepsilon_H = \sup_{\mathbf{u}} \|\widehat{y}_H - P_\psi(q_H)\|$, planner optimization error by $\varepsilon_{\mathrm{opt}}$, and planning regret by $\mathrm{Reg}_H = J_Q(\widehat{\mathbf{u}}) - \min_{\mathbf{u}} J_Q(\mathbf{u})$.

To study training dynamics, we use a diagonalizable population model

$$x_{i,t+1} = \rho_i x_{i,t} + B_i^\top u_t + \xi_{i,t}, \qquad \mathbb{E}[x_{i,t}^2] = \sigma_i^2, \qquad \mathbb{E}[\xi_{i,t}^2] = \omega_i^2. \tag{5}$$

For mode $i$, $s_i = (\sigma_i w_i)^2$ is its representation energy, normalized by $\sum_i s_i = 1$, $e_i = p_i - \rho_i$ its predictor mismatch, and $\nu_i = \omega_i^2/\sigma_i^2$ its innovation cost. Modes are partitioned into decision modes $D$ and nuisance modes $N$. For a fixed directional bias $\gamma_i$, define $\bar{\nu}_i = \nu_i - \lambda\gamma_i$ and selection score $\theta_i = \bar{\nu}_i + e_i^2$. Encoder and predictor learning rates are $\eta_E, \eta_P > 0$.

For one decision mode $d$ and one nuisance mode $n$, write $p = s_d$, $r = p/(1-p)$, $y = e_d^2$, $z = e_n^2$, and $a = \bar{\nu}_n - \bar{\nu}_d$. Thus $a > 0$ favors the decision mode once both predictors are fitted. A score intervention $v(t) \ge 0$ reduces the decision score by $v(t)$ without changing the predictor update; $U \ge 0$ denotes an imposed strength limit. We use a prescribed threshold $\tau \in (0, 1)$ to require $p \ge \tau$, a modal retention condition distinct from the planning metric above. Subscript 0 denotes the state at intervention onset, and $(x)_+ = \max\{x, 0\}$.

## 4 HOW PREDICTIVE TRAINING SELECTS THE WRONG STATE

The failure we study is not ordinary representation collapse. A latent can remain diverse and perfectly predictable while discarding the state needed for control. The following construction makes this distinction exact.

**Theorem 4.1** (Marginal non-identifiability)**.** *Let $R(E) = \mathfrak{R}(E_{\#}\mu)$ be any marginal-only constraint admitting a nonatomic target distribution $\rho$ on a standard Borel space. There exists a fully observed deterministic controlled process and two encoder–predictor pairs $(E_D, F_D)$ and $(E_N, F_N)$ such that both have zero prediction loss, both induce $\rho$, and hence both attain the same value of $R$, while only $E_D$ is decision-sufficient.*

**Remark (what non-collapse certifies).** Variance, rank, or marginal matching can certify that a latent encodes something, but not that it encodes distinctions that change a decision. A model may therefore look healthy under prediction and anti-collapse diagnostics while being unusable for planning. Evaluation must test action-conditioned distinctions rather than latent activity alone.

The theorem shows that an endpoint objective cannot by itself tell the optimizer which non-collapsed solution to select. We next examine that selection during training. In the modal model of Section 3, let

$$\mathcal{J}(s,e) = \sum_i s_i\theta_i, \qquad \theta_i = \bar{\nu}_i + e_i^2, \qquad \sum_i s_i = 1.$$

**Theorem 4.2** (General coupled mode selection)**.** *For any finite number of modes, projected encoder gradient flow and ordinary predictor gradient flow satisfy*

$$\dot{s}_i = -4\eta_E(\theta_i - \bar{\theta})s_i, \qquad \dot{e}_i = -2\eta_P s_i e_i, \qquad \bar{\theta} = \sum_j s_j\theta_j. \tag{6}$$

*Thus a low-score mode gains representation, and representation determines how quickly its predictor improves. In particular,*

$$e_i(t) = e_i(0)e^{-2\eta_P\int_0^t s_i(\tau)d\tau}, \quad \log\frac{s_i(t)}{s_j(t)} = \log\frac{s_i(0)}{s_j(0)} - 4\eta_E\int_0^t[\theta_i(\tau) - \theta_j(\tau)]d\tau. \tag{7}$$

*More generally, for any partition into decision modes $D$ and nuisance modes $N$,*

$$\frac{d}{dt}\log\frac{S_D}{S_N} = -4\eta_E(\bar{\theta}_D - \bar{\theta}_N), \tag{8}$$

*where $S_A = \sum_{i\in A} s_i$ and $\bar{\theta}_A = S_A^{-1}\sum_{i\in A} s_i\theta_i$.*

**Remark (why the wrong feature wins).** A mode that gains energy early learns a better predictor, which lowers its score and attracts still more energy. A suppressed mode receives little predictor training; its mismatch persists and continues to suppress it. Joint training therefore rewards an early predictive advantage twice. This explains how successful optimization can improve prediction by committing to an easy nuisance.

**Proposition 1** (Capacity-limited subspace selection)**.** *Let a rank-$d$ encoder optimize a time-independent symmetric selection operator with ordered eigenvalues $\theta_1 \le \cdots \le \theta_m$. If $\theta_d < \theta_{d+1}$ and the initialization has full-rank projection onto the bottom-$d$ eigenspace, its representation subspace converges to that eigenspace at rate $O(e^{-2\eta(\theta_{d+1}-\theta_d)t})$. If the necessary decision subspace contains $r \le d$ eigenmodes, all of them survive exactly when*

$$\max_{i\in D}\theta_i < \theta_{N,(d-r+1)}, \tag{9}$$

*in the absence of ties.*

**Remark.** Increasing latent dimension allows more nuisance modes to coexist with the decision state, but it does not change their ordering. Once predictable nuisances occupy all available slots ahead of a necessary decision mode, non-collapse and additional training cannot bring that mode back; the training objective must change its relative score.

To characterize when this feedback becomes irreversible, consider one necessary decision mode $d$ and one nuisance mode $n$. Write

$$r = \frac{s_d}{s_n}, \qquad y = e_d^2, \qquad z = e_n^2, \qquad a = \bar{\nu}_n - \bar{\nu}_d.$$

Let $v(t)$ be the relative score bonus induced by a training constraint, with $v > 0$ favoring the decision mode. The coupled flow becomes

$$\frac{d}{dt}\log r = 4\eta_E[a + v(t) + z - y], \qquad \dot{y} = -4\eta_P\frac{r}{1+r}y, \qquad \dot{z} = -4\eta_P\frac{1}{1+r}z. \tag{10}$$

The first identity already gives an operational criterion: having ratio $r(T) \ge R > r_0$ at time $T$ is equivalent to accumulating selection margin

$$4\eta_E\int_0^T[a + v(t) + z(t) - y(t)]dt \ge \log\frac{R}{r_0}. \tag{11}$$

Intervention therefore depends on both direction and timing, rather than its final loss value alone.

This intervention has a concrete meaning for an implemented constraint. Let $\mathcal{R} = \lambda R(c_d, c_n)$ denote the complete weighted score-only term in the encoder objective. In the interior $c_d c_n \neq 0$, its effective directional contribution is

$$v_{\mathcal{R}}(t) = -\frac{1}{2}\left(\frac{\partial_{c_d}\mathcal{R}}{c_d} - \frac{\partial_{c_n}\mathcal{R}}{c_n}\right). \tag{12}$$

The common radial term created by the representation-budget projection cancels from this difference. Thus the bound $U$ controls the regularizer's measured gradient in the useful direction, not its nominal loss coefficient. Eligibility for the exact boundary additionally requires bounded $v_{\mathcal{R}}$ and no direct change to the predictor flow. DQF also trains a planning head and predictor branches, so it is a theory-guided witness rather than an implementation of the exact controller.

The mature-nuisance regime admits a sharper result. Suppose the nuisance predictor is already fitted, so $z_0 = 0$, while the decision predictor is not. For a bounded score intervention $v(t) \leq U$, define

$$b = a + U, \qquad c = \frac{\eta_P}{\eta_E}, \qquad \Psi_b(y) = y - b - b\log(y/b).$$

**Theorem 4.3** (Exact recoverability boundary)**.** *Assume $r_0 > 0$, $z_0 = 0$, $b > 0$, $y_0 > b$, and $v \in L^1_{\mathrm{loc}}$ with $0 \leq v(t) \leq U$ almost everywhere. Define*

$$\Phi_0 = c\log(1 + r_0) - \Psi_b(y_0). \tag{13}$$

*Then $\Phi_0 = 0$ is the exact boundary for bounded score intervention. If $\Phi_0 \leq 0$, every admissible intervention yields $r(t) \to 0$ and $y(t) \to y_\infty \geq b > 0$. If $\Phi_0 > 0$, the maximal intervention $v(t) = U$ yields $r(t) \to \infty$ and $y(t) \to 0$.*

**Remark (from modal share to readability).** Modal share is a mechanism rather than an evaluation metric. A simple stable measurement bridge makes the connection explicit: if a normalized decision variable is observed through $\widetilde{z}_d = \sqrt{s_d}x_d + \xi$, with independent noise $\xi \sim \mathcal{N}(0, \sigma^2)$, the population-optimal linear probe has $R^2 = s_d/(s_d + \sigma^2)$. Hence the theorem's limit $s_d \to 0$ forces stable decision readability to vanish whenever measurement noise is nonzero. Our experiments report held-out probes and planning, rather than equating a small internal coefficient with information loss by definition.

## 5 Protecting Decision Information During Joint Training

The recovery boundary concerns information that is already being lost. Prevention asks how much intervention keeps a decision mode available while its predictor learns. Fix a retained-share threshold $\tau \in (0, 1)$ and start from $p_0 \geq \tau$. Under the two-mode flow,

$$\dot{p} = 4\eta_E p(1-p)(a + z - y + v), \qquad \dot{y} = -4\eta_P p y, \qquad \dot{z} = -4\eta_P(1-p)z. \tag{14}$$

The score correction $v \geq 0$ acts on encoder selection, leaving predictor learning unchanged.

**Theorem 5.1** (Minimal protection and predictor learning)**.** *Let $a \in \mathbb{R}$, $\eta_E, \eta_P > 0$, $p_0 \in [\tau, 1)$, and $y_0, z_0 \geq 0$. The boundary correction*

$$v_\tau(p, y, z) = \begin{cases} 0, & p > \tau, \\ (y - z - a)_+, & p = \tau \end{cases} \tag{15}$$

*defines a unique absolutely continuous protected trajectory, with the flow interpreted almost everywhere. It preserves $p(t) \geq \tau$ and gives*

$$y(t) = y_0 \exp\left(-4\eta_P \int_0^t p(s)\, ds\right) \leq y_0 e^{-4\eta_P \tau t}. \tag{16}$$

*At each boundary state, $(y - z - a)_+$ is the smallest nonnegative score correction satisfying $\dot{p} \geq 0$. With a cap $U$, this correction is feasible exactly when $(y - z - a)_+ \leq U$ at every visited boundary state; $U \geq (y_0 - a)_+$ is sufficient for the whole trajectory.*

**Remark (what protection buys).** The threshold guarantees learning opportunity rather than a prescribed latent distribution. A larger retained share gives a faster guaranteed reduction of predictor mismatch. The required correction depends on the current selection disadvantage, not on a fixed penalty weight. This is instantaneous minimality at the same state: the theorem specifies how much directional correction is needed to prevent further loss.

Together, Theorem 4.3 and Theorem 5.1 give a state-dependent design rule: intervene before a bounded correction can no longer recover the decision mode, preserve a chosen decision share, and correct only its outward drift at the threshold. This is a single continuous training process, not a pretraining–finetuning schedule.

**From retained information to planning.** The threshold $\tau$ specifies retained modal energy. Planning additionally depends on how that information is represented and predicted. The faithfulness condition in Section 3 provides the corresponding endpoint guarantee.

**Theorem 5.2** (Geometry-to-planning guarantee)**.** *Suppose the planning map $P = P_\psi$ is $(\alpha, \beta)$-faithful on the reachable set, the $H$-step rollout error is at most $\varepsilon_H$, and the planner returns $\widehat{\mathbf{u}}$ within $\varepsilon_{\mathrm{opt}}$ of the minimum of $\widehat{J}$. For a true optimal sequence $\mathbf{u}^\star$,*

$$J_Q(\widehat{\mathbf{u}}) \le \frac{\beta}{\alpha} J_Q(\mathbf{u}^\star) + \frac{2\varepsilon_H + \varepsilon_{\mathrm{opt}}}{\alpha}. \tag{17}$$

**Remark (what planning needs).** Protecting information enables its predictor to learn; usable geometry lets a planner exploit it. The bound separates representation distortion from rollout and optimization error, making planning success the downstream test rather than treating latent non-collapse as a substitute.

**Action-conditioned constraints.** Our controlled experiments use a decision-quotient feasibility constraint (DQF) to expose action-responsive information to a planning head $P_\psi$. Let $U_c$ denote centered action-consequence coordinates, $Y_c$ observed endpoint head outputs, and $\widehat{Y}_c$ predicted endpoint head outputs. For a nuisance-equivalent pair $h \sim_H h'$, write

$$\ell_{\mathrm{eq}} = \|Y_c - U_c\|_F^2, \qquad \ell_{\mathrm{fib}} = \|P_\psi(h) - P_\psi(h')\|^2, \qquad \ell_{\mathrm{tr}} = \|\widehat{Y}_c - \mathrm{sg}(Y_c)\|_F^2.$$

The evaluated constraint is

$$R_{\mathrm{DQF}} = (\ell_{\mathrm{eq}} - \epsilon_{\mathrm{eq}}^2)_+^2 + (\ell_{\mathrm{fib}} - \epsilon_{\mathrm{fib}}^2)_+^2 + (\ell_{\mathrm{tr}} - \epsilon_{\mathrm{tr}}^2)_+^2. \tag{18}$$

The coordinate certificate is given in Appendix 8.7. On a finite connected constraint graph, zero hinge penalty implies a per-edge coordinate error no larger than $\epsilon_G = \max\{\sqrt{2}\,\epsilon_{\mathrm{eq}}, \epsilon_{\mathrm{fib}}\}$, and hence a global error proportional to the graph diameter. This statement applies only to the sampled connected support and is relative to the chosen consequence coordinate $\chi$; planning faithfulness additionally requires $\chi$ itself to preserve decision distance. DQF tests action-conditioned feature protection; Eq. equation 15 specifies the separate state-dependent update required for the modal protection guarantee. Implementing that update in a neural model requires a decision-share estimate and a correction that enforces the boundary condition. A tolerance hinge alone does not enforce it.

## 6 Experiments

The experiments connect the training mechanism to its practical consequence: whether the learned state supports planning. We use the LeWM encoder and action-conditioned latent predictor (Maes et al., 2026). SIGReg and DQF (Eq. equation 18) share the architecture, predictive objective, optimizer, data, and training and CEM planning budgets. Main results use five seeds, 3,000 updates, and 100 held-out goals per seed. We report prediction loss, held-out linear probes, endpoint distance, and success at distance below $0.20$.

**Separating decision state from predictable nuisance.** Figure 1 shows the two controlled tasks. The cyan agent responds to actions, whereas the Wallpaper texture and Distractor's orange object are predictable but irrelevant to reaching the goal. Current and goal images have different nuisance, so matching their appearance is not enough. This construction instantiates the decision–nuisance

distinction in Section 3: the model must retain action-responsive state despite competing visual variation.

DQF therefore receives richer relational supervision than SIGReg. The comparison below asks whether such decision-directed protection can alter representation selection and planning. We reserve method-level equivalence claims for controls that expose both objectives to the same pairs, branches, and planning head.

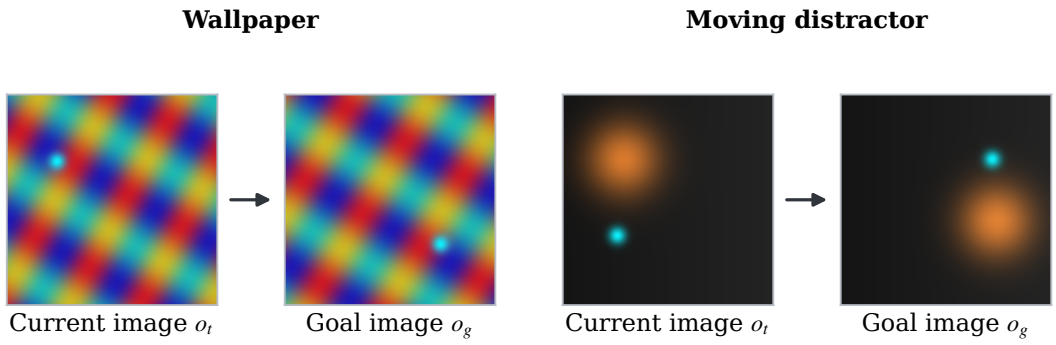


Figure 1: Controlled visual goal-reaching environments. A planner must move the cyan agent from the current image to its position in the goal image using continuous two-dimensional actions. Current and goal observations contain different predictable nuisance: a full-frame texture in Wallpaper and an independently moving orange object in Distractor.

**Prediction improves without decision sufficiency.** Table 1 tests the distinction established by Theorem 4.1. On Wallpaper, SIGReg achieves lower prediction loss but almost never reaches the goal; DQF retains a decision-readable planning quotient and substantially reduces endpoint distance. Distractor gives the same planning benefit with less severe baseline failure. Prediction loss therefore does not rank planning quality. The probes diagnose readable state, while planning tests whether the model can use it—the separate downstream requirement in Theorem 5.2.

Table 1: Prediction, representation, and planning after joint training (five seeds; mean $\pm$ standard deviation). Decision probes use SIGReg's latent and DQF's planning quotient, respectively.

| Task | Method | Pred. loss $\downarrow$ | Decision $R^2$ $\uparrow$ | Success $\uparrow$ | Distance $\downarrow$ |
|---|---|---|---|---|---|
| Wallpaper | SIGReg | $0.0055 \pm 0.0008$ | $-0.071 \pm 0.030$ | $0.004 \pm 0.009$ | $1.146 \pm 0.090$ |
| Wallpaper | DQF | $0.0072 \pm 0.0004$ | $0.971 \pm 0.002$ | $0.442 \pm 0.145$ | $0.239 \pm 0.036$ |
| Distractor | SIGReg | $0.0425 \pm 0.0029$ | $0.652 \pm 0.281$ | $0.148 \pm 0.086$ | $0.549 \pm 0.126$ |
| Distractor | DQF | $0.0589 \pm 0.0014$ | $0.975 \pm 0.002$ | $0.458 \pm 0.194$ | $0.263 \pm 0.086$ |

**What is selected during training?** Figure 2 distinguishes a training-induced failure from an uninformative endpoint comparison. Prediction improves under both objectives, but SIGReg's nuisance readability increases while decision readability remains near chance. DQF's quotient instead develops decision readability during training. This is the selection pattern predicted by Theorem 4.2: reducing predictive error need not favor the state required for control, and an action-directed constraint can change which distinctions become represented. Linear probes measure readability rather than modal energy; the trajectories reveal the direction of selection without identifying the flow's parameters. Their practical message is to monitor decision-relevant distinctions while prediction improves, not only the final loss.

**When does targeted protection help?** Figure 3 asks whether the failure depends systematically on the task's competing variation. We vary background amplitude while keeping controlled dynamics and evaluation fixed. The horizontal axis measures nuisance-induced pixel variation relative to one-step action-induced variation, making severity observable from intervention pairs rather than an arbitrary task index. As this ratio increases, SIGReg's decision–nuisance probe gap becomes negative and planning degrades despite low prediction loss. DQF becomes beneficial in this regime,

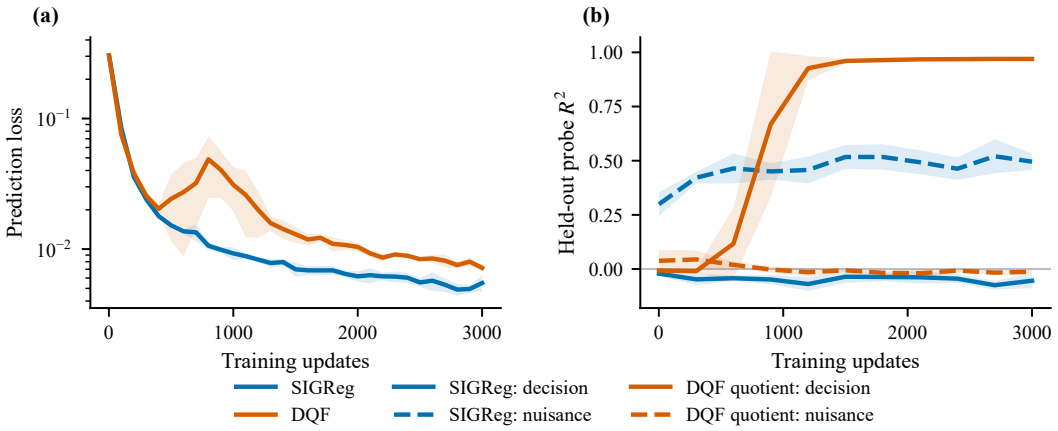


Figure 2: Training dynamics on Wallpaper (five seeds; bands show one standard deviation). Prediction improves under both objectives. SIGReg's nuisance probe improves while its decision probe remains near chance; DQF's quotient becomes decision-readable with low nuisance readability.

while SIGReg remains competitive at weak nuisance. The sweep supports the competition account in Section 4 and identifies when targeted protection is useful; its empirical transition is not the exact state-space separatrix of Theorem 4.3.

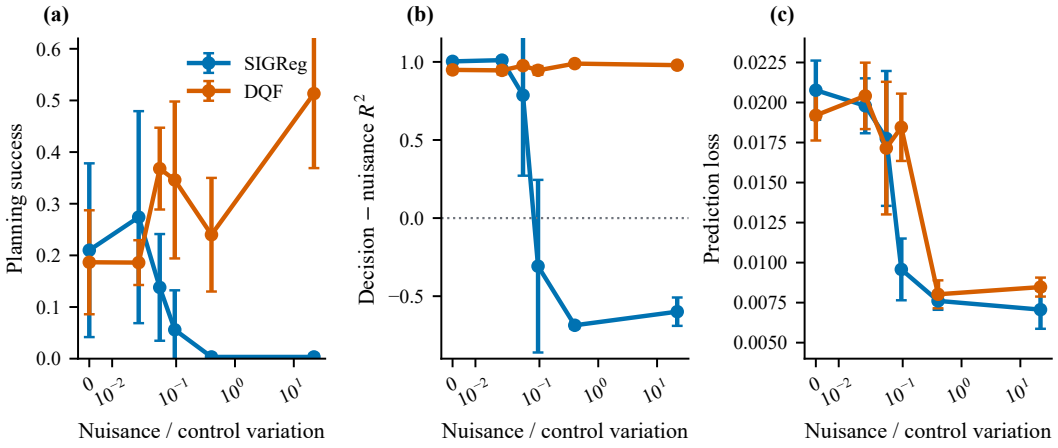


Figure 3: Nuisance–control phase sweep. The horizontal axis measures nuisance-induced pixel variation relative to a one-step action. Settings have at least three seeds; bars show one standard deviation. SIGReg's decision–nuisance probe gap becomes negative despite low prediction loss, whereas DQF maintains a positive gap.

**Finite-step learning reproduces the recovery boundary.** Figure 4 directly tests Theorem 4.3. Rather than integrating the theorem's ODE, we optimize the underlying diagonal encoder and predictor by finite-step projected gradient descent from a grid of initial representation shares and predictor mismatches. The learned systems recover precisely on the side predicted by the analytic separatrix. Across nine intervention-strength and learning-rate settings (7,569 initial states), pointwise agreement is at least 99.88% and is 100% outside a narrow margin around the boundary. This is an exact test of the tractable population model, not a claim that nonlinear pixel models share the same separatrix. It establishes the sharp mechanism; the visual tasks test whether its qualitative consequences survive in a learned world model.

**Protection must favor the action-responsive direction.** Figure 5 removes DQF's action-consequence coordinate while retaining its other terms. Endpoint distance worsens on both tasks;

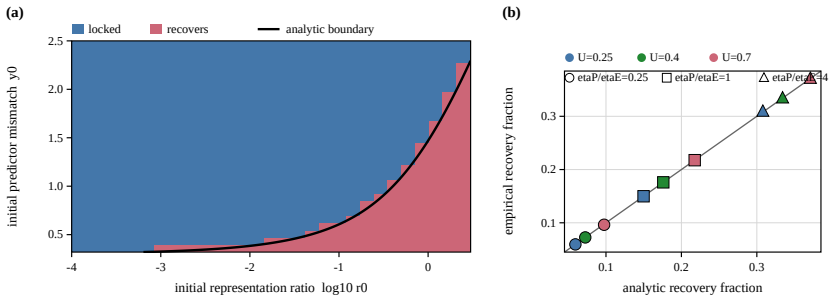


Figure 4: Finite-step validation of the exact recovery boundary. **(a)** Outcomes of 841 independently initialized learned encoder–predictor systems at $U = 0.4$ and $\eta_P/\eta_E = 1$; the black curve is the analytic separatrix. **(b)** Empirical versus analytic recovery fractions for three intervention strengths (color) and three learning-rate ratios (marker). The diagonal is equality.

on Wallpaper, decision readability disappears even though prediction loss improves. This is partial evidence for the directional requirement in Theorem 5.1; because the unconstrained head can itself degenerate, a variance-matched no-coordinate control is needed to isolate direction from head non-collapse. A post-hoc head on a frozen SIGReg encoder also fails to recover planning, whereas a DQF-trained encoder retains part of its gain with full-latent planning (Appendix 9). Together, these comparisons place part of the benefit in representation formation, not only the planning head. DQF tests this principle empirically; it does not implement the theorem's exact boundary correction.

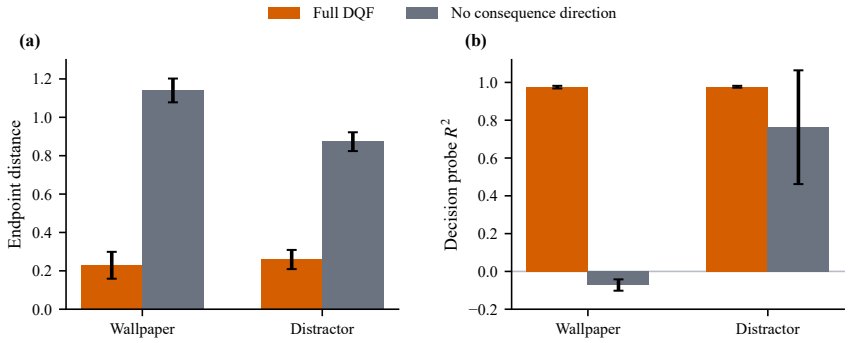


Figure 5: Direction ablation (three paired seeds). Removing the action-consequence coordinate worsens endpoint distance on both tasks and eliminates decision readability on Wallpaper.

These controlled experiments connect predictive feature selection to planning and identify when decision-directed intervention helps.

## 7 Conclusion

A world model's predictive success depends on which information its training dynamics choose to retain. We characterize a feedback in which reduced decision representation slows predictor learning, and the resulting mismatch further suppresses that representation. In the modal model, this feedback yields a quantitative recovery boundary under bounded intervention. Preserving a chosen decision share guarantees continued predictor learning, and the minimal boundary correction identifies the direction and instantaneous strength required to prevent further loss. Controlled visual planning experiments show how accurate prediction coexists with decision-information loss and how action-conditioned constraints alter feature selection. Together, these results give constraint design a concrete training-time objective: protect the information needed for decisions before predictive training suppresses it. Translating the modal criterion into measurable neural training signals remains an important direction for adaptive world-model constraints.

## REFERENCES


Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Assran_Self-Supervised_Learning_From_Images_With_a_Joint-Embedding_Predictive_Architecture_CVPR_2023_paper.html.

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

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.

Aditya Sharad Golatkar, Alessandro Achille, and Stefano Soatto. Time matters in regularizing deep networks: Weight decay and data augmentation affect early learning dynamics, matter little near convergence. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/87784eca6b0dea1dff92478fb786b401-Abstract.html.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent—a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2555–2565. PMLR, 2019.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025. doi: 10.1038/s41586-025-08744-2. URL https://www.nature.com/articles/s41586-025-08744-2.

Nick Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/cf73d57b6dcda32b293df7c2d5341f49-Abstract-Conference.html.

JangHyeon Lee, Philipe Dias, Yao-Yi Chiang, and Dalton Lunga. Understanding the learning phases in self-supervised learning via critical periods. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/86ddf3543ad437d71c37e510f41b1a53-Abstract-Conference.html.

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. *arXiv preprint arXiv:2407.03475*, 2024.

Chang Liu, Fei Suo, Yanzhou Jin, Zeyu Ping, Yusuke Iwasawa, Yutaka Matsuo, and Yaonan Zhu. Temporally centered SIGReg improves LeWorldModel representations for robot policy learning. *arXiv preprint arXiv:2607.26924*, 2026. URL https://arxiv.org/abs/2607.26924v3.

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

Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C. Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/0987b8b338d6c90bbedd8631bc499221-Abstract.html.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.6120.

Vlad Sobal, Jyothir S V, Siddhartha Jalagam, Nicolas Carion, Kyunghyun Cho, and Yann LeCun. Joint embedding predictive architectures focus on slow features. *arXiv preprint arXiv:2211.10831*, 2022.

Kaizhen Tan, Xin Xu, Siru Tao, Yixiao Li, Hanzhe Hong, Yang Feng, and Heqing Du. What can latent world models know? physical parameter identifiability in multimodal predictive representations. *arXiv preprint arXiv:2607.27017*, 2026. URL https://arxiv.org/abs/2607.27017v3.

Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10268–10278. PMLR, 2021. URL https://proceedings.mlr.press/v139/tian21a.html.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320. PMLR, 2021. URL https://proceedings.mlr.press/v139/zbontar21a.html.

# 8 Proofs and Additional Theory

## 8.1 Marginal non-identifiability

**Proof of Theorem 4.1.** Let the observed state be $x = (q, n) \in \mathbb{S}^1 \times \mathbb{S}^1$, let the action space be $\mathbb{S}^1$, and consider

$$q_{t+1} = q_t + u_t \pmod 1, \qquad n_{t+1} = n_t.$$

Take the training-state distribution to be uniform on the torus. Since $\rho$ is nonatomic on a standard Borel space, the isomorphism theorem for standard probability spaces gives full-measure subsets related by a measure-preserving Borel isomorphism $\psi : \mathbb{S}^1 \to \mathcal{Z}$. Define

$$E_D(q, n) = \psi(q), \qquad E_N(q, n) = \psi(n).$$

Both pushforwards equal $\rho$, so every marginal-only constraint assigns them the same value. On the full-measure set where $\psi^{-1}$ is defined, set

$$F_D(z, u) = \psi\big(\psi^{-1}(z) + u \pmod 1\big), \qquad F_N(z, u) = z.$$

The two latent dynamics are exact almost surely, hence both prediction losses vanish.

For any two states $(q, n)$ and $(q', n)$ with $q \neq q'$, $E_N(q, n) = E_N(q', n)$. Their controlled future kernels differ because, under the same action sequence, their terminal controllable coordinates differ by $q - q'$. Thus no map from $E_N$ can recover the decision quotient. In contrast, $\psi^{-1} \circ E_D = q$ almost surely, so $E_D$ is decision-sufficient. This proves the claim. □

### 8.2 General modal dynamics

We first derive the flow for an arbitrary finite collection of modes. Let $c = (c_1, \dots, c_m)$ be the standardized encoder coefficients, constrained by $\|c\|_2 = 1$, and write $s_i = c_i^2$. At fixed predictor parameters,

$$\mathcal{J}(c, e) = \sum_{i=1}^{m} c_i^2 \theta_i, \qquad \theta_i = \bar{\nu}_i + e_i^2.$$

**Proof of Theorem 4.2.** For projected encoder gradient flow with rate $\eta_E$ and predictor gradient flow with rate $\eta_P$,

$$\dot{s}_i = -4\eta_E(\theta_i - \bar{\theta})s_i, \qquad \bar{\theta} = \sum_{k=1}^{m} s_k \theta_k, \tag{19}$$

$$\dot{e}_i = -2\eta_P s_i e_i. \tag{20}$$

For every pair with positive initial energy,

$$e_i(t) = e_i(0) \exp\left[-2\eta_P \int_0^t s_i(\tau) d\tau\right], \tag{21}$$

$$\log \frac{s_i(t)}{s_j(t)} = \log \frac{s_i(0)}{s_j(0)} - 4\eta_E \int_0^t [\theta_i(\tau) - \theta_j(\tau)] d\tau. \tag{22}$$

The Euclidean encoder gradient is $\nabla_c \mathcal{J} = 2\,\mathrm{diag}(\theta)c$. Projection onto the tangent space of the unit sphere gives

$$\dot{c} = -2\eta_E(I - cc^\top)\,\mathrm{diag}(\theta)c.$$

Because $c^\top \mathrm{diag}(\theta)c = \bar{\theta}$, its $i$-th component is $\dot{c}_i = -2\eta_E(\theta_i - \bar{\theta})c_i$. Differentiating $s_i = c_i^2$ proves equation 19. The predictor derivative is $\partial\mathcal{J}/\partial e_i = 2s_i e_i$, which proves equation 20. Equation equation 21 follows by integrating $\dot{e}_i/e_i = -2\eta_P s_i$; it remains valid when $e_i(0) = 0$. Taking the difference of $\dot{s}_i/s_i = -4\eta_E(\theta_i - \bar{\theta})$ for two modes and integrating proves equation 22. Positivity is preserved by the exponential form of the solution.

The same identity applies to groups of modes. For a partition $D \cup N = \{1, \dots, m\}$, define

$$S_D = \sum_{i \in D} s_i, \qquad \bar{\theta}_D = \frac{1}{S_D} \sum_{i \in D} s_i \theta_i,$$

and define $S_N, \bar{\theta}_N$ analogously. Summing equation 19 within each group gives

$$\frac{d}{dt} \log \frac{S_D}{S_N} = -4\eta_E(\bar{\theta}_D - \bar{\theta}_N). \tag{23}$$

This proves the group identity and completes the proof. □

**Proof of Proposition 1.** Let $C = C^\top \in \mathbb{R}^{m \times m}$ be a time-independent effective selection operator, and let $X \in \mathbb{R}^{m \times d}$ satisfy $X^\top X = I_d$. For $\mathcal{J}(X) = \mathrm{tr}(X^\top C X)$, Grassmann gradient flow is

$$\dot{X} = -2\eta(I - XX^\top)CX. \tag{24}$$

Writing $\Pi = XX^\top$,

$$\dot{\Pi} = -2\eta(C\Pi + \Pi C - 2\Pi C\Pi), \tag{25}$$

$$\frac{d}{dt}\,\mathrm{tr}(C\Pi) = -4\eta\|(I - \Pi)C\Pi\|_F^2 \le 0. \tag{26}$$

If the ordered eigenvalues satisfy $\theta_d < \theta_{d+1}$ and the initial subspace has full-rank projection onto the bottom-$d$ eigenspace, then $\Pi(t)$ converges to that eigenspace, with principal-angle error

$$O\left(e^{-2\eta(\theta_{d+1} - \theta_d)t}\right).$$

Projecting $2CX$ onto the horizontal tangent space of the Grassmann manifold yields equation 24. Differentiating $\Pi = XX^\top$ and substituting equation 24 gives equation 25. Cyclicity of trace then gives

$$\frac{d}{dt}\operatorname{tr}(C\Pi) = -4\eta \operatorname{tr}(\Pi C(I-\Pi)C\Pi) = -4\eta\|(I-\Pi)C\Pi\|_F^2.$$

To identify the limit, let $C = V\operatorname{diag}(\theta_1,\dots,\theta_m)V^\top$ and $Y(t) = e^{-2\eta Ct}X(0)$. Orthonormalizing the columns of $Y(t)$ produces the same subspace trajectory as equation 24. In the eigenbasis of $C$, every component above the bottom-$d$ eigenspace is attenuated relative to its slowest retained component by at least $e^{-2\eta(\theta_{d+1}-\theta_d)t}$. The full-rank projection assumption keeps all bottom-$d$ directions present, so the principal angles converge at the stated rate.

Suppose the eigenvectors of $C$ are the modes and the necessary decision set has size $r \le d$. If nuisance scores in increasing order are $\theta_{N,(1)} \le \theta_{N,(2)} \le \cdots$, then, in the absence of ties, every decision mode belongs to the selected rank-$d$ subspace exactly when

$$\max_{i\in D}\theta_i < \theta_{N,(d-r+1)}. \tag{27}$$

At most $d-r$ nuisance modes may precede the worst-ranked decision mode. This counting condition is necessary and sufficient and proves equation 9. □

### 8.3 Cumulative recovery and the exact lock-in boundary

For the two-mode reduction, let

$$r = \frac{s_d}{s_n}, \qquad y = e_d^2, \qquad z = e_n^2, \qquad a = \bar\nu_n - \bar\nu_d.$$

Since $s_d = r/(1+r)$ and $s_n = 1/(1+r)$, Theorem 4.2 gives

$$\frac{d}{dt}\log r = 4\eta_E[a + v(t) + z - y], \tag{28}$$

$$\dot y = -4\eta_P\frac{r}{1+r}y, \qquad \dot z = -4\eta_P\frac{1}{1+r}z. \tag{29}$$

For completeness, let $\mathcal{R} = \lambda R(c_d, c_n)$ be the complete weighted differentiable score-only term in the encoder loss. Before projection its gradient contribution is $-\eta_E(\partial_{c_d}\mathcal{R}, \partial_{c_n}\mathcal{R})$. On the interior $c_d c_n \neq 0$, the unit-sphere projection subtracts a common radial scalar times $(c_d, c_n)$. Dividing each coordinate flow by its coordinate and subtracting therefore cancels that scalar, giving

$$\frac{d}{dt}\log\frac{s_d}{s_n}\bigg|_{\mathcal{R}} = -2\eta_E\left(\frac{\partial_{c_d}\mathcal{R}}{c_d} - \frac{\partial_{c_n}\mathcal{R}}{c_n}\right) = 4\eta_E v_{\mathcal{R}}.$$

This proves Eq. equation 12. It also explains why the theorem bounds effective directional gradient rather than a raw regularization weight.

Integrating equation 28 proves equation 11. In particular, if the integrand is bounded below by $\delta > 0$ on $[0,T]$, then

$$r(T) \ge r_0 e^{4\eta_E\delta T},$$

so every target $R > r_0$ is reached by

$$T \ge \frac{1}{4\eta_E\delta}\log\frac{R}{r_0}.$$

**Proof of Theorem 4.3.** Reset the intervention time to zero. Under $z_0 = 0$, uniqueness of equation 29 gives $z(t) = 0$. First take the maximal admissible intervention $v(t) = U$, and set

$$b = a + U > 0, \qquad c = \frac{\eta_P}{\eta_E}.$$

Along this flow,

$$\mathcal{H}(r,y) = b\log y - y + c\log(1+r)$$

is conserved. Indeed,

$$\frac{d}{dt}(b\log y - y) = -4\eta_P \frac{r}{1+r}(b-y),$$

while

$$\frac{d}{dt}\big[c\log(1+r)\big] = 4\eta_P \frac{r}{1+r}(b-y).$$

The function $h_b(y) = b\log y - y$ has derivative $b/y - 1$ and a unique maximum at $y = b$. For $y_0 > b$, conservation at a possible first crossing $y = b$ requires

$$c\log(1+r_b) = c\log(1+r_0) - \Psi_b(y_0), \qquad \Psi_b(y_0) = y_0 - b - b\log\frac{y_0}{b}.$$

If $\Phi_0 = c\log(1+r_0) - \Psi_b(y_0) < 0$, no $r_b \geq 0$ satisfies this equality. The trajectory remains in $y > b$, where $r$ decreases. Its limit must have $r_\infty = 0$: a positive limit would make $s_d$ bounded below, force $y$ below $b$, and contradict the absence of a crossing. Conservation then determines a limit $y_\infty > b$.

If $\Phi_0 = 0$, the conserved level reaches the maximum of $h_b$ only at $r = 0$. Hence $r(t) \to 0$ and $y(t) \to b$.

If $\Phi_0 > 0$, the crossing ratio is

$$r_b = \exp(\Phi_0/c) - 1 > 0.$$

Before the first crossing, conservation and maximality of $h_b(b)$ imply $r(t) \geq r_b$. Therefore

$$-\frac{d}{dt}\log y(t) = 4\eta_P \frac{r(t)}{1+r(t)} \geq 4\eta_P \frac{r_b}{1+r_b},$$

so $y$ reaches $b$ in finite time. After crossing, $b - y > 0$ and $r$ grows. If $\int_0^\infty s_d(t)dt < \infty$, then $y$ would converge to a positive value below $b$, making the eventual growth rate of $\log r$ positive. This would imply $r \to \infty$, hence $s_d \to 1$, contradicting the finite integral. Thus $\int_0^\infty s_d(t)dt = \infty$, which gives $y(t) \to 0$. Equation equation 28 then gives $r(t) \to \infty$.

It remains to compare $v = U$ with every admissible $v(t) \leq U$. Define

$$A(t) = \int_0^t \frac{r(\tau)}{1+r(\tau)}d\tau, \qquad y(t) = y_0 e^{-4\eta_P A(t)}.$$

The state $(r, A)$ satisfies

$$\dot{r} = 4\eta_E\big[a + v(t) - y_0 e^{-4\eta_P A}\big]\, r, \qquad \dot{A} = \frac{r}{1+r}.$$

Both cross derivatives are nonnegative:

$$\partial_A \dot{r} = 16\eta_E\eta_P y_0 e^{-4\eta_P A} r \geq 0, \qquad \partial_r \dot{A} = (1+r)^{-2} > 0.$$

The vector field is therefore cooperative and its first component is monotone in $v$. The comparison theorem for cooperative Carathéodory systems yields $r_v(t) \leq r_U(t)$ and $A_v(t) \leq A_U(t)$ from equal initial states. Consequently, if maximal intervention lies on or below the boundary, no admissible intervention recovers the decision mode; above the boundary, maximal intervention recovers it. □

The lock-in region persists when the nuisance predictor is not fully fitted. Since $z(t) \leq z_0$, set

$$b_+ = a + U + z_0.$$

Comparison with the more favorable auxiliary system that keeps $z(t) \equiv z_0$ shows that

$$y_0 > b_+, \qquad c\log(1+r_0) < y_0 - b_+ - b_+\log\frac{y_0}{b_+} \tag{30}$$

is sufficient for failure of every $v(t) \leq U$. These strict inequalities define a nonempty open set: choose $y_0 = 2b_+$ and

$$0 < r_0 < \exp\left(\frac{b_+(1-\log 2)}{c}\right) - 1.$$

Within an arbitrary finite-mode system, any invariant decision–nuisance face obeys the same equations, while equation 23 gives the corresponding exact aggregate selection law away from such a face.

**Stable-readout bridge.** Let $x_d \sim \mathcal{N}(0,1)$ and let a noisy observed decision coordinate be $\widetilde{z}_d = \sqrt{s_d}x_d + \xi$, where $\xi \sim \mathcal{N}(0,\sigma^2)$ is independent. The population-optimal linear predictor of $x_d$ from $\widetilde{z}_d$ has

$$R^2_{\mathrm{opt}} = \frac{\mathrm{Cov}(x_d,\widetilde{z}_d)^2}{\mathrm{Var}(x_d)\,\mathrm{Var}(\widetilde{z}_d)} = \frac{s_d}{s_d+\sigma^2}.$$

Thus for every fixed $\sigma^2 > 0$, the lock-in limit $s_d \to 0$ implies $R^2_{\mathrm{opt}} \to 0$. This bridge is intentionally minimal: without noise, norm control, or another stability condition, an arbitrarily small nonzero coefficient need not imply information-theoretic loss.

### 8.4 Minimal protection

**Proof of Theorem 5.1.** At $p = \tau$, the factor $4\eta_E\tau(1-\tau)$ is positive. Thus $\dot{p} \geq 0$ is equivalent to $v \geq y - z - a$. Among nonnegative corrections the unique minimum is $(y-z-a)_+$; in the interior the least instantaneous correction is zero.

To establish a well-defined protected trajectory, set $\ell = \log(p/(1-p))$ and $c_\tau = \log(\tau/(1-\tau))$. For a continuous path $x$ starting at $\ell_0 \geq c_\tau$, define

$$K_x(t) = \sup_{0\leq s\leq t}(c_\tau - x(s))_+, \qquad \Gamma(x) = x + K_x.$$

Then $\Gamma(x) \geq c_\tau$, and $K_x$ increases only at the boundary. Moreover,

$$\|K_x - K_{\widetilde{x}}\|_\infty \leq \|x-\widetilde{x}\|_\infty, \qquad \|\Gamma(x)-\Gamma(\widetilde{x})\|_\infty \leq 2\|x-\widetilde{x}\|_\infty.$$

For a trial path $\ell$, let

$$p_\ell = (1+e^{-\ell})^{-1}, \quad y_\ell(t) = y_0 e^{-4\eta_P\int_0^t p_\ell(s)ds}, \quad z_\ell(t) = z_0 e^{-4\eta_P\int_0^t (1-p_\ell(s))ds},$$

and set

$$x_\ell(t) = \ell_0 + 4\eta_E\int_0^t (a + z_\ell(s) - y_\ell(s))ds.$$

The logistic function is $1/4$-Lipschitz, so on $[0,T]$,

$$\|y_\ell - y_{\widetilde{\ell}}\|_\infty \leq \eta_P y_0 T\|\ell - \widetilde{\ell}\|_\infty, \qquad \|z_\ell - z_{\widetilde{\ell}}\|_\infty \leq \eta_P z_0 T\|\ell-\widetilde{\ell}\|_\infty.$$

Consequently, $\ell \mapsto \Gamma(x_\ell)$ has Lipschitz constant at most $8\eta_E\eta_P(y_0+z_0)T^2$. It is a contraction for sufficiently small $T$, yielding a unique local fixed point. Since $0 \leq y \leq y_0$, $0 \leq z \leq z_0$, and the driving path has bounded derivative, the construction extends to every finite interval.

For the fixed point, $K_x$ is locally absolutely continuous. Its support property and the vanishing derivative of an absolutely continuous function on a level set imply, almost everywhere,

$$\dot{K}_x = \begin{cases} 0, & \ell > c_\tau, \\ [-4\eta_E(a+z-y)]_+, & \ell = c_\tau. \end{cases}$$

Thus $v = \dot{K}_x/(4\eta_E)$ realizes Eq. equation 15, with no impulses, and $p \geq \tau$. Conversely any protected solution of this rule has the same reflection representation, so uniqueness follows from the fixed point. Integrating the predictor equation gives

$$y(t) = y_0 e^{-4\eta_P\int_0^t p(s)ds} \leq y_0 e^{-4\eta_P\tau t}.$$

At a visited boundary state, the cap is feasible exactly when $(y-z-a)_+ \leq U$. Since $(y-z-a)_+ \leq (y_0-a)_+$, the stated uniform bound is sufficient. □

### 8.5 Metric and curvature evolution

For completeness, spatially varying selection can also be described through the first variation of scalar curvature. Let $g(t)$ be a smooth family of Riemannian metrics, $h = \partial_t g$, $\mathrm{Ric}$ its Ricci tensor, and $\mathrm{Scal}$ its scalar curvature. With $\Delta = \nabla^i\nabla_i$,

$$\partial_t \mathrm{Scal} = -\langle h, \mathrm{Ric}\rangle_g + \nabla^i\nabla^j h_{ij} - \Delta\,\mathrm{tr}_g\, h. \tag{31}$$

To verify it, differentiating the Levi–Civita connection gives

$$\partial_t \Gamma^k_{ij} = \frac{1}{2} g^{k\ell} \left( \nabla_i h_{j\ell} + \nabla_j h_{i\ell} - \nabla_\ell h_{ij} \right).$$

Substitution into $\mathrm{Ric}_{ij} = \partial_k \Gamma^k_{ij} - \partial_j \Gamma^k_{ik} + \Gamma^k_{k\ell}\Gamma^\ell_{ij} - \Gamma^k_{j\ell}\Gamma^\ell_{ik}$, followed by covariant contraction, yields

$$g^{ij} \partial_t \mathrm{Ric}_{ij} = \nabla^i \nabla^j h_{ij} - \Delta \, \mathrm{tr}_g \, h.$$

The derivative of the inverse metric is $\partial_t g^{ij} = -h^{ij}$; contracting $\mathrm{Scal} = g^{ij}\mathrm{Ric}_{ij}$ then proves equation 31.

For $g = G_P = e^{2\varphi} G_Q$ in dimension two, $h = 2\dot{\varphi} g$, $\mathrm{Ric} = K_P g$, and $\mathrm{Scal} = 2K_P$. Substitution into equation 31 gives

$$\partial_t K_P = -2\dot{\varphi} K_P - \Delta_{G_P} \dot{\varphi}.$$

The two-dimensional conformal Laplacian satisfies $\Delta_{G_P} = e^{-2\varphi} \Delta_{G_Q}$, hence

$$\partial_t K_P = -2\dot{\varphi} K_P - e^{-2\varphi} \Delta_{G_Q} \dot{\varphi}. \tag{32}$$

### 8.6 Planning guarantees

**Proof of Theorem 5.2.** For every candidate action sequence $\mathbf{u}$, faithfulness and the rollout bound give

$$\alpha J_Q(\mathbf{u}) \le \| P(q_H(\mathbf{u}; h)) - P(q_g) \| \le \widehat{J}(\mathbf{u}) + \varepsilon_H.$$

Applying this to $\widehat{\mathbf{u}}$, then using planner optimality, the rollout bound for $\mathbf{u}^\star$, and the upper geometric inequality gives

$$\begin{aligned} \alpha J_Q(\widehat{\mathbf{u}}) &\le \widehat{J}(\widehat{\mathbf{u}}) + \varepsilon_H \\ &\le \widehat{J}(\mathbf{u}^\star) + \varepsilon_{\mathrm{opt}} + \varepsilon_H \\ &\le \| P(q_H(\mathbf{u}^\star; h)) - P(q_g) \| + 2\varepsilon_H + \varepsilon_{\mathrm{opt}} \\ &\le \beta J_Q(\mathbf{u}^\star) + 2\varepsilon_H + \varepsilon_{\mathrm{opt}}. \end{aligned}$$

Division by $\alpha > 0$ proves equation 17. The proof uses only the metric inequalities on the reachable set and therefore applies to arbitrary decision metric spaces and action-sequence classes. □

**Lemma A.3 (common and quotient rollout errors).** For $M$ candidate action sequences, let $z_m$ and $\widehat{z}_m$ be their true encoded and predicted endpoints. With

$$\bar{z} = \frac{1}{M} \sum_m z_m, \qquad \bar{\widehat{z}} = \frac{1}{M} \sum_m \widehat{z}_m,$$

define

$$e_0 = \bar{\widehat{z}} - \bar{z}, \qquad e^Q_m = (\widehat{z}_m - \bar{\widehat{z}}) - (z_m - \bar{z}).$$

Then

$$\widehat{z}_m - z_m = e_0 + e^Q_m, \qquad \max_m \| \widehat{z}_m - z_m \| \le \| e_0 \| + \max_m \| e^Q_m \|.$$

**Proof.** Adding and subtracting the two means yields the equality:

$$\widehat{z}_m - z_m = (\bar{\widehat{z}} - \bar{z}) + (\widehat{z}_m - \bar{\widehat{z}}) - (z_m - \bar{z}).$$

The bound follows from the triangle inequality and maximization over $m$. □

The residual $\ell_{\mathrm{tr}}$ controls centered differences $e^Q_m$. Absolute goal-distance planning also requires control of the common error $e_0$, since shifting all predicted endpoints without shifting the goal can change their ranking.

**Corollary A.4 (estimated decision metric).** Suppose smooth metrics satisfy

$$(1-\delta)G_Q \preceq \widehat{G}_Q \preceq (1+\delta)G_Q, \qquad 0 \le \delta < 1,$$

and

$$\alpha^2 \widehat{G}_Q \preceq G_P \preceq \beta^2 \widehat{G}_Q.$$

Then

$$\alpha\sqrt{1-\delta}\, d_Q \le d_P \le \beta\sqrt{1+\delta}\, d_Q,$$

where $d_P$ is the intrinsic length distance induced by $G_P$. These constants apply to Theorem 5.2 when the planner uses this distance, or when the same bounds are separately established for its Euclidean endpoint distance.

**Proof.** Transitivity of the Loewner order gives

$$\alpha^2(1-\delta)G_Q \preceq G_P \preceq \beta^2(1+\delta)G_Q.$$

Applying the quadratic-form bounds to every path tangent, taking square roots, integrating path length, and then taking the infimum over paths proves the distance inequalities. □

### 8.7 Decision-quotient feasibility

**Proposition 2** (Finite-graph coordinate certificate)**.** *Let $\chi : Q \to \mathbb{R}^k$ be the consequence coordinate used in $U_c$, and let $G = (V, E)$ be the finite constraint graph whose edges join co-branch endpoints or nuisance-equivalent histories. Write $\delta_v = P_\psi(v) - \chi(\pi(v))$. If $G$ is connected with graph diameter at most $L$, and every edge satisfies $\|\delta_v - \delta_w\| \le \epsilon$, then for any anchor $v_0$, with $c = \delta_{v_0}$,*

$$\max_{v \in V} \|P_\psi(v) - \chi(\pi(v)) - c\| \le L\epsilon.$$

*Consequently, the distance distortion between any two certified vertices is at most $L\epsilon$. At $\epsilon = 0$, this recovers $P_\psi = \chi \circ \pi + c$ exactly on the connected support. If the transport residual also vanishes, predicted pairwise endpoint differences equal the observed differences on the certified branches.*

**Proof of Proposition 2.** Choose a shortest path $v_0, v_1, \ldots, v_m = v$. Since $m \le L$, telescoping and the triangle inequality give

$$\|\delta_v - \delta_{v_0}\| \le \sum_{j=1}^{m} \|\delta_{v_j} - \delta_{v_{j-1}}\| \le L\epsilon.$$

Taking $c = \delta_{v_0}$ proves the first claim. For any $v, w \in V$, apply the same telescoping argument directly to a shortest $v$-to-$w$ path, whose length is at most $L$. The reverse triangle inequality then yields

$$\big|\|P_\psi(v) - P_\psi(w)\| - \|\chi(\pi(v)) - \chi(\pi(w))\|\big| \le \|\delta_v - \delta_w\| \le L\epsilon.$$

The exact statement follows by setting $\epsilon = 0$. Finally, zero transport residual gives $\widehat{Y}_c = Y_c$ on every observed branch, so predicted and observed pairwise consequences agree there. □

To relate the proposition to Eq. equation 18, write the centered coordinate residual of one branch as $C\Delta$. If $\|C\Delta\|_F \le \epsilon_{\mathrm{eq}}$, then for any two endpoints $i, j$ in that branch,

$$\|\delta_i - \delta_j\| = \|(C\Delta)_i - (C\Delta)_j\| \le \sqrt{2}\,\|C\Delta\|_F \le \sqrt{2}\,\epsilon_{\mathrm{eq}}.$$

For a nuisance-equivalent edge, $\chi \circ \pi$ agrees at both endpoints, so $\ell_{\mathrm{fib}} \le \epsilon_{\mathrm{fib}}^2$ gives $\|\delta_i - \delta_j\| \le \epsilon_{\mathrm{fib}}$. Thus zero hinge penalty certifies the proposition with $\epsilon_G = \max\{\sqrt{2}\,\epsilon_{\mathrm{eq}}, \epsilon_{\mathrm{fib}}\}$ on the observed graph. This expression uses the unaveraged Frobenius residual in Eq. equation 18; a mean-reduced implementation must restore the corresponding square-root cardinality factor. The same calculation bounds pairwise centered transport error by $\sqrt{2}\,\epsilon_{\mathrm{tr}}$.

The certificate is relative to $\chi$. It does not by itself prove decision sufficiency or the multiplicative faithfulness assumed by Theorem 5.2. Those conclusions require $\chi$ to be injective or faithful on the certified quotient states; converting the additive $L\epsilon_G$ error to a multiplicative lower bound also requires positive separation at the scale being certified.

**Inactive within the feasible band.** For $f_\epsilon(\ell) = (\ell - \epsilon^2)_+^2$, both $f_\epsilon(\ell)$ and its derivative vanish whenever $\ell \le \epsilon^2$. Each term of $R_{\mathrm{DQF}}$ therefore has zero value and zero parameter gradient inside its feasible band. This is a property of the penalty, not invariance of that band under the full training dynamics. Proposition 2 controls only the finite connected constraint graph; extending it beyond sampled support requires separate coverage and smoothness assumptions.

**Selection margin.** Under the modal flow, let $\theta_i^R$ denote the complete score after adding the constraint. Equation equation 22 gives

$$\log \frac{s_i(T)}{s_j(T)} = \log \frac{s_i(0)}{s_j(0)} + 4\eta_E \int_0^T [\theta_j^R(t) - \theta_i^R(t)]dt.$$

Thus the target ratio $s_i(T)/s_j(T) \ge R$ is equivalent to a cumulative score margin of at least $\log(Rs_j(0)/s_i(0))$. For the two-mode reduction this is Eq. equation 11.

## 9 Additional Experimental Diagnostics

On Wallpaper, SIGReg prediction loss decreases from $0.304$ to $0.005$ while nuisance probe $R^2$ increases from $0.299$ to $0.495$ and decision readability remains near chance. Under DQF, decision information emerges between updates 300 and 600 and reaches $R^2 = 0.932$ at update 1,200.

**Fixed-budget intervention timing.** Figure 6 crosses the DQF weight with the update at which the constraint is activated. Strong protection applied from the beginning best preserves decision readability and gives the lowest planning error; delaying the same intervention reduces both benefits. All runs stop after 3,000 updates, so a later start also receives fewer corrective updates. The figure therefore provides a practical fixed-budget timing test, not evidence for the exact infinite-time recovery boundary in Theorem 4.3.

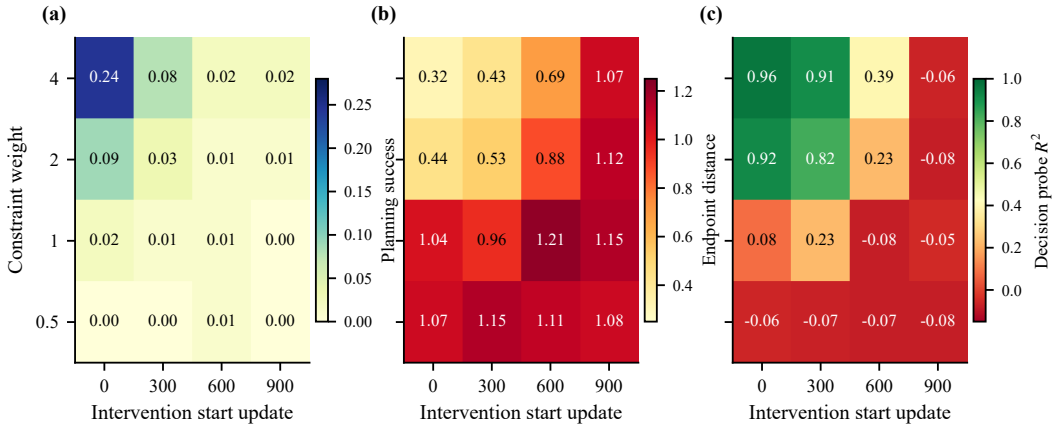


Figure 6: Constraint strength and activation time under a common 3,000-update budget (three seeds per cell). **(a)** Planning success (higher is better), **(b)** endpoint distance (lower is better), and **(c)** decision-probe $R^2$ (higher is better). Later starts receive fewer constrained updates; the comparison tests fixed-budget training rather than the exact analytic separatrix.

In the direction ablation, removing the action-consequence coordinate improves Wallpaper prediction loss from $0.00849$ to $0.00494$, but reduces decision $R^2$ from $0.975$ to $-0.072$. The learned-boundary experiment evaluates 841 initial states for each of nine strength and learning-rate settings. The minimum full-grid agreement with the analytic classification is $99.88\%$; after excluding the declared narrow boundary margin, every setting has $100\%$ agreement.

A post-hoc quotient head fitted to a frozen SIGReg encoder gives planning success $0.002$ on Wallpaper and $0.128$ on Distractor. Full-latent planning with a DQF-trained encoder gives $0.210$ and $0.242$, compared with $0.442$ and $0.458$ using its quotient head. These diagnostics separate changes in encoded information from changes in the planner's readout.

### 9.1 DETAILED EXPERIMENTAL RESULTS

Tables 2–6 report the numerical values underlying the main experimental comparisons. Unless noted otherwise, standard deviations are sample standard deviations across seeds. For DQF, decision and nuisance probes in these tables are measured on the planning quotient; SIGReg probes use its predictive latent.

Table 2: Per-seed results for the main comparison. Each planning entry uses 100 independently sampled goals.

| Task | Method | Seed | Pred. loss | Decision $R^2$ | Nuisance $R^2$ | Success | Distance |
|---|---|---|---|---|---|---|---|
| Wallpaper | SIGReg | 0 | 0.00561 | -0.057 | 0.448 | 0.00 | 1.246 |
| Wallpaper | SIGReg | 1 | 0.00461 | -0.035 | 0.511 | 0.00 | 1.237 |
| Wallpaper | SIGReg | 2 | 0.00486 | -0.117 | 0.424 | 0.02 | 1.044 |
| Wallpaper | SIGReg | 3 | 0.00650 | -0.068 | 0.569 | 0.00 | 1.096 |
| Wallpaper | SIGReg | 4 | 0.00590 | -0.075 | 0.435 | 0.00 | 1.108 |
| Wallpaper | DQF | 0 | 0.00671 | 0.973 | -0.023 | 0.49 | 0.232 |
| Wallpaper | DQF | 1 | 0.00750 | 0.969 | -0.027 | 0.36 | 0.271 |
| Wallpaper | DQF | 2 | 0.00754 | 0.970 | -0.010 | 0.23 | 0.282 |
| Wallpaper | DQF | 3 | 0.00682 | 0.970 | -0.014 | 0.58 | 0.207 |
| Wallpaper | DQF | 4 | 0.00736 | 0.972 | -0.012 | 0.55 | 0.205 |
| Distractor | SIGReg | 0 | 0.04638 | 0.377 | 0.925 | 0.09 | 0.564 |
| Distractor | SIGReg | 1 | 0.04434 | 0.495 | 0.911 | 0.10 | 0.653 |
| Distractor | SIGReg | 2 | 0.04129 | 0.478 | 0.921 | 0.07 | 0.690 |
| Distractor | SIGReg | 3 | 0.03874 | 0.977 | 0.927 | 0.26 | 0.417 |
| Distractor | SIGReg | 4 | 0.04185 | 0.932 | 0.923 | 0.22 | 0.423 |
| Distractor | DQF | 0 | 0.02988 | 0.977 | 0.002 | 0.53 | 0.211 |
| Distractor | DQF | 1 | 0.02790 | 0.978 | -0.021 | 0.58 | 0.202 |
| Distractor | DQF | 2 | 0.02929 | 0.975 | 0.009 | 0.67 | 0.189 |
| Distractor | DQF | 3 | 0.03025 | 0.972 | -0.012 | 0.22 | 0.371 |
| Distractor | DQF | 4 | 0.02699 | 0.973 | -0.009 | 0.29 | 0.342 |

Table 3: Wallpaper training trajectories underlying Figure 2. Entries are means over five seeds. DQF probe values use its planning quotient.

| Method | Update | Pred. loss | Decision $R^2$ | Nuisance $R^2$ |
|---|---|---|---|---|
| SIGReg | 1 | 0.3043 | -0.021 | 0.299 |
| SIGReg | 300 | 0.0242 | -0.048 | 0.421 |
| SIGReg | 600 | 0.0137 | -0.042 | 0.464 |
| SIGReg | 1200 | 0.0083 | -0.069 | 0.457 |
| SIGReg | 3000 | 0.0055 | -0.054 | 0.495 |
| DQF | 1 | 0.3043 | -0.007 | 0.038 |
| DQF | 300 | 0.0256 | -0.010 | 0.044 |
| DQF | 600 | 0.0273 | 0.117 | 0.019 |
| DQF | 1200 | 0.0200 | 0.927 | -0.014 |
| DQF | 3000 | 0.0072 | 0.970 | -0.013 |

Table 4: Nuisance–control sweep underlying Figure 3. We include every setting evaluated with at least three seeds. The variation ratio is nuisance-induced pixel MSE divided by one-step action-induced pixel MSE. The probe gap is decision $R^2$ minus nuisance $R^2$ in the representation used for planning.

| Nuisance strength | Variation ratio | Method | Seeds | Probe gap | Success | Pred. loss |
|---|---|---|---|---|---|---|
| 0.00 | 0.000 | SIGReg | 3 | $1.003 \pm 0.001$ | $0.210 \pm 0.168$ | $0.0208 \pm 0.0018$ |
| 0.00 | 0.000 | DQF | 3 | $0.949 \pm 0.020$ | $0.187 \pm 0.101$ | $0.0192 \pm 0.0016$ |
| 0.04 | 0.023 | SIGReg | 5 | $1.011 \pm 0.016$ | $0.274 \pm 0.205$ | $0.0198 \pm 0.0017$ |
| 0.04 | 0.023 | DQF | 5 | $0.946 \pm 0.025$ | $0.186 \pm 0.043$ | $0.0204 \pm 0.0021$ |
| 0.06 | 0.053 | SIGReg | 5 | $0.787 \pm 0.515$ | $0.138 \pm 0.103$ | $0.0178 \pm 0.0042$ |
| 0.06 | 0.053 | DQF | 5 | $0.976 \pm 0.011$ | $0.368 \pm 0.079$ | $0.0172 \pm 0.0041$ |
| 0.08 | 0.095 | SIGReg | 5 | $-0.308 \pm 0.553$ | $0.056 \pm 0.076$ | $0.0096 \pm 0.0019$ |
| 0.08 | 0.095 | DQF | 5 | $0.946 \pm 0.027$ | $0.346 \pm 0.152$ | $0.0185 \pm 0.0021$ |
| 0.16 | 0.399 | SIGReg | 3 | $-0.686 \pm 0.020$ | $0.003 \pm 0.006$ | $0.0076 \pm 0.0006$ |
| 0.16 | 0.399 | DQF | 3 | $0.989 \pm 0.011$ | $0.240 \pm 0.110$ | $0.0080 \pm 0.0009$ |
| 1.00 | 22.008 | SIGReg | 3 | $-0.599 \pm 0.091$ | $0.003 \pm 0.006$ | $0.0071 \pm 0.0012$ |
| 1.00 | 22.008 | DQF | 3 | $0.979 \pm 0.001$ | $0.513 \pm 0.144$ | $0.0085 \pm 0.0006$ |

Table 5: Learned-system validation underlying Figure 4. Each row contains 841 initial states. "Agreement" is pointwise agreement between finite-step optimization and the analytic recovery classification; "away" excludes the fixed $|\Phi_0| < 0.02$ margin.

| $U$ | $\eta_P/\eta_E$ | Empirical | Analytic | Agreement / away |
|---|---|---|---|---|
| 0.25 | 0.25 | 0.0595 | 0.0595 | 1.0000 / 1.0000 |
| 0.25 | 1.00 | 0.1498 | 0.1498 | 1.0000 / 1.0000 |
| 0.25 | 4.00 | 0.3092 | 0.3080 | 0.9988 / 1.0000 |
| 0.40 | 0.25 | 0.0725 | 0.0725 | 1.0000 / 1.0000 |
| 0.40 | 1.00 | 0.1760 | 0.1760 | 1.0000 / 1.0000 |
| 0.40 | 4.00 | 0.3341 | 0.3341 | 1.0000 / 1.0000 |
| 0.70 | 0.25 | 0.0963 | 0.0975 | 0.9988 / 1.0000 |
| 0.70 | 1.00 | 0.2176 | 0.2176 | 1.0000 / 1.0000 |
| 0.70 | 4.00 | 0.3710 | 0.3710 | 1.0000 / 1.0000 |

Table 6: Direction ablation underlying Figure 5 (three paired seeds; mean $\pm$ standard deviation). The no-transport condition is included for completeness but does not establish that transport is necessary on these short-horizon tasks.

| Task | Variant | Success | Distance | Decision $R^2$ | Pred. loss |
|---|---|---|---|---|---|
| Wallpaper | Full DQF | $0.460 \pm 0.207$ | $0.229 \pm 0.070$ | $0.975 \pm 0.007$ | 0.00849 |
| Wallpaper | No coordinate direction | $0.010 \pm 0.017$ | $1.140 \pm 0.062$ | $-0.072 \pm 0.030$ | 0.00494 |
| Wallpaper | No transport | $0.377 \pm 0.151$ | $0.257 \pm 0.052$ | $0.977 \pm 0.003$ | 0.00703 |
| Distractor | Full DQF | $0.437 \pm 0.129$ | $0.259 \pm 0.050$ | $0.977 \pm 0.005$ | 0.02949 |
| Distractor | No coordinate direction | $0.037 \pm 0.021$ | $0.873 \pm 0.049$ | $0.763 \pm 0.301$ | 0.04299 |
| Distractor | No transport | $0.457 \pm 0.229$ | $0.246 \pm 0.082$ | $0.978 \pm 0.001$ | 0.02854 |

### 9.2 Training configuration and compute

All comparisons use the same LeWM backbone and optimizer within an experiment; only the stated constraint or sweep variable changes. Table 7 gives the shared configuration. The main comparison, training trajectories, and matched-representation controls ran on one NVIDIA A100-SXM4 80GB. The nuisance sweep and direction ablations ran on NVIDIA RTX 4090 24GB GPUs. Runs used PyTorch 2.10 with CUDA 12.8 or 12.9 and automatic mixed precision in bfloat16. The deterministic boundary validation uses vectorized float64 population gradients on CPU. No distributed training was used.

Table 7: Training and evaluation configuration. The main comparison, nuisance sweep, and direction ablation use 100 planning goals per run. The boundary validation is deterministic and therefore does not use random seeds or planning goals.

| Component | Configuration |
|---|---|
| Observation and history | RGB images of size $64\times64$; three-frame history; continuous two-dimensional actions. |
| Encoder | LeWM ViT-Tiny, patch size 8, trained from scratch; projection MLP $192 \to 768 \to 16$. |
| Latent predictor | Six transformer blocks, hidden width 192, eight attention heads, MLP width 768, head dimension 24, dropout 0.1; latent and action-embedding dimension 16. |
| Optimization | AdamW; 3,000 updates; batch size 128; learning rate $10^{-4}$ for encoder and predictor; weight decay $10^{-3}$; bfloat16 mixed precision. |
| SIGReg | Weight 0.09, 17 knots, and 256 projections. |
| DQF | Overall weight 2.0; coordinate, fiber, and transport weights all 1.0; all feasibility tolerances 0.10; one-step consequence branches. |
| Diagnostics | Linear probes every 300 updates using 512 held-out samples; main seeds $\{0,1,2,3,4\}$; ablation and intervention seeds $\{0,1,2\}$. |
| Planning | CEM horizon 5, 384 samples, 48 elites, four iterations; 100 held-out goals for the main results; success means endpoint distance below 0.20. |
| Compute | One NVIDIA A100-SXM4 80GB or one NVIDIA RTX 4090 24GB per run; PyTorch 2.10; no distributed training. |