# CAPTURE IS NOT PROPAGATION: AUDITING ACTION EFFECTS IN LATENT WORLD MODELS

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

## ABSTRACT

A latent world model can encode a task variable while predicting the wrong effect of acting on it. We separate these capabilities with a capture-gated audit: readouts fitted to real data must capture the current query and resolve executed action effects before they are used to score imagined effects, so that propagation failures are not confused with readout failures. Three LeWorldModel checkpoints on Cheetah pass both gates, yet their imagined five-step effects have $R^2$ from $-6.33$ to $-3.36$, well below the zero-effect baseline. The decoded effects are misaligned and amplified even though the complete feature changes align closely with the real ones. Severity varies across Walker, Finger, and other model groups, and the same directional failure appears in independently trained TD-MPC2 runs. On held-out Cheetah action sets, disagreement between model-predicted effects and an environment-fitted reference ranks regret among the candidate actions, so open-loop effect error is visible at the decision level. The same experiments show the limit of any such score: it compares candidate actions with one another, while most of the realized loss comes from the candidate set itself, which no relative score can detect. Evaluating a world model for planning therefore requires testing action effects directly, after capture, and against a no-intervention reference.

## 1 INTRODUCTION

A planner needs a world model to encode decision-relevant variables and predict how actions change them (Sutton, 1991; Ha & Schmidhuber, 2018). Readability from an encoded observation does not establish correct imagined action effects; neither one-step accuracy nor realized return separates these capabilities (Lambert et al., 2020; Vafa et al., 2024).

We audit *planning-time intervention fidelity*: the change in a declared task query under matched environment actions and the model's native open-loop transition. Frozen root and effect readouts test current-query capture and executed-effect resolution; only then does the gated audit compare imagined effects with executed effects, no effect, and executed-endpoint decoding. Unresolved real contrasts stop that audit.

On Cheetah, three frozen LeWorldModel checkpoints (Maes et al., 2026) pass both prerequisites, but their imagined five-step effects have $R^2$ between $-6.33$ and $-3.36$. All three intervals lie below zero, and the imagined endpoints are worse than the oracle under the same readout. Although the complete predicted feature changes align closely with the real changes, the decoded task effects are misaligned and amplified.

Severe failures also occur on Walker, vary across LeWorldModel Finger seeds, and are absent from the PreJEPA group, although every candidate that passes both gates is worse than the oracle. A matched-shell protocol finds the same directional deficit in three independently trained TD-MPC2 Cheetah runs at two radii (Hansen et al., 2024). In a descriptive Finger comparison, the same readout accurately decodes large intervention effects from observed endpoints but fails on five-step imagined endpoints. Small-effect decoding generalized poorly in the preceding gate test, so this comparison uses the large radius.

On held-out sets of six Cheetah actions, the discrepancy score ranks regret within the candidate set in all six run-radius evaluations, with clear gains over a capacity-matched random-feature control in three. Two post-hoc analyses then ask what the score does not capture. On Cheetah, most of the



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.

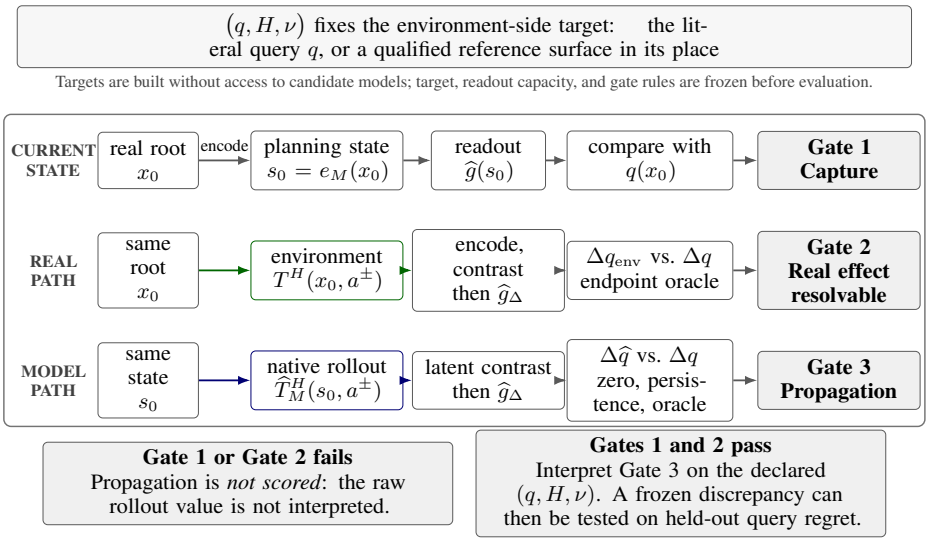



Figure 1: **Capture-gated matched-intervention audit.** The frozen readout must capture the query at real roots and resolve the matched effect at executed endpoints before imagined propagation is scored. Passing capture is a prerequisite, not evidence of propagation. Query regret is evaluated separately on held-out cases.

loss against a benchmark that includes no intervention comes from the candidate set itself, and the score does not track that component. On Finger, a signed-effect rule selects a small subset of actions with positive average effect relative to no intervention, but its paired difference from always selecting spans zero.

## 2 A Capture-Gated Audit of Matched Action Effects

### 2.1 Setup

Let $x_0$ be a real initial (root) state, $q(x)$ a declared scalar or vector task query, $H$ a physical horizon, and $\nu$ a distribution over roots and coupled action pairs $(a^+, a^-)$. How the pair is coupled is part of the declared support: the pulse studies contrast matched actuator pulses executed from a shared root, and the shell studies contrast actions drawn from a fixed set of action sequences (Section 2.3). Root support names the distribution from which initial states are drawn: task-policy roots come from the trained policy's visitation distribution, and environment-random roots come from the environment's reset distribution. Model $M$ exposes an encoder $e_M$, a planner-facing state $s_0 = e_M(x_0)$, and a native open-loop $H$-step transition $\widehat{T}_M^H$. The native interface fixes how the $H$ physical steps are advanced: as one macro-step prediction consuming the full action block (LeWorldModel, PreJEPA) or as an autoregressive composition of native steps (TD-MPC2); the audit scores each model under its own declared interface. Two readouts are trained only on real data, cross-fit while keeping dependent samples together, and frozen before the candidate is evaluated: a root readout $\widehat{g}$ from encoded states to the query, and an effect readout $\widehat{g}_\Delta$ from endpoint representation contrasts to the executed effect. Neither receives action identity. All norms use coordinate scales estimated on development data.

The confirmatory audits proceed in gate order (Figure 1), preventing a missing target, or one the readout cannot resolve, from being misread as a transition result.

**Gate 1: root capture.** On held-out real roots, query capture is the centered coefficient of determination

$$R^2_{\text{cap}} = 1 - \frac{\sum_i \|\widehat{g}(e_M(x_0^i)) - q(x_0^i)\|^2}{\sum_i \|q(x_0^i) - \bar{q}\|^2}. \tag{1}$$

Pulse capture requires $R^2_{\text{cap}} > 0$ for each query coordinate and a one-sided, 256-draw family-preserving permutation $p \le 0.05$, with conventions frozen before scoring. Appendix C gives the shell additions. The readout is then frozen.

**Gate 2: real-effect resolution.** For matched actions $a^+$ and $a^-$ from a shared root, let $x_H^+$ and $x_H^-$ be their executed environment endpoints. The true and readout-resolved effects are

$$\Delta q = q(x_H^+) - q(x_H^-), \qquad \Delta q_{\text{env}} = \widehat{g}_\Delta\big(e_M(x_H^+) - e_M(x_H^-)\big). \tag{2}$$

The effect readout consumes one differenced representation and is fit to executed effects; decoding the two endpoint levels with $\widehat{g}$ and differencing provides a separate sensitivity check where reported. The score

$$R^2_{\text{real}} = 1 - \frac{\sum_i \|\Delta q^i_{\text{env}} - \Delta q^i\|^2}{\sum_i \|\Delta q^i\|^2} \tag{3}$$

is an uncentered, zero-referenced coefficient of determination: its baseline is predicting no effect, and no mean is subtracted (unlike Equation 1). Resolution passes under the same frozen convention as capture: $R^2_{\text{real}} > 0$ and permutation $p \le 0.05$. A cell that fails this gate has an unresolved effect and no interpretable propagation score.

**Gate 3: imagined propagation.** The same root and actions are advanced through the model, $\widehat{s}_H^{\pm} = \widehat{T}_M^H(s_0, a^{\pm})$, producing $\Delta\widehat{q} = \widehat{g}_\Delta(\widehat{s}_H^+ - \widehat{s}_H^-)$. Replacing the executed effect in Equation 3 gives

$$R^2_{\text{prop}} = 1 - \frac{\sum_i \|\Delta\widehat{q}^i - \Delta q^i\|^2}{\sum_i \|\Delta q^i\|^2}. \tag{4}$$

A score of $R^2_{\text{prop}} < 0$ means that the imagined action effect is worse than predicting no effect. The *oracle deficit* compares errors paired within environment families: with $\mathrm{RMSE}_f(\cdot)$ the root-mean-squared error of an effect estimate against $\Delta q$ within family $f$, over the $F$ evaluation families,

$$\Delta_{\text{orc}} = \frac{1}{F}\sum_{f=1}^{F}\Big[\mathrm{RMSE}_f\big(\Delta\widehat{q}\big) - \mathrm{RMSE}_f\big(\Delta q_{\text{env}}\big)\Big], \tag{5}$$

with whole-family bootstrap intervals. A positive deficit means the imagined endpoint reads out worse than the executed endpoint under the same frozen effect readout; figure axes label this quantity as the RMSE gap. Controls include persistence, action null, wrong actions, label and feature shuffles, and readout capacity. Persistence repeats the root representation at the endpoint, so nothing evolves over the horizon; action-null instead rolls the model forward under all-zero actions, separating a frozen state from a removed intervention.

### 2.2 Design, inference, and protocol boundaries

Environment families are independent reset identities that supply repeated roots and action contrasts. The environment and model receive the same physical action sequence after interface mapping. Development, validation, readout training, and evaluation use disjoint family identities; evaluation does not set targets, readout capacity, scales, thresholds, candidates, or action support. The complete environment family is the uncertainty unit. Seeds and candidates are reported separately.

The primary studies use matched actuator pulses. The Cheetah shell protocol instead uses fixed-radius action sets and a model-independent response surface fitted and qualified on disjoint environment families, and per-candidate root and effect readouts on the planning latent. This fitted reference surface approximates responses on the tested shells. The descriptive Finger comparison uses centered effect $R^2$ and a pooled RMSE difference between observed and imagined endpoints (Appendix E).

The finite-horizon error decomposition that motivates the gate order separates root capture, restricted transition error, environment approximation residuals, and horizon-query residuals (Appendix A). A synthetic ten-case suite verifies the stopping logic and required invariances (Appendix A.1).

### 2.3 From shell discrepancy to held-out query regret

For each initial state $c$, the fixed-radius test compares a set $\mathcal{A}_c$ of candidate action sequences. Let $d_M(c,a)$ and $d_C(c,a)$ be the model and qualified reference-surface query displacements. For the six actions compared here,

$$s(c) = \max_{\{a,a'\}\subset\mathcal{A}_c} \left|[d_M(c,a) - d_M(c,a')] - [d_C(c,a) - d_C(c,a')]\right|. \tag{6}$$

The maximum ranges over all 15 unordered action pairs, allowing the score to be frozen before any action is chosen or executed. Including the model-preferred and surface-preferred actions among these pairs gives the regret bound below. The resulting bound can far exceed realized regret (Section 5). Its scale and positive bias depend on the number of actions, so scores are not compared across action counts $K$.

Held-out *query regret within the candidate set* compares the environment query achieved by the model-selected action with that of the best candidate action, under the frozen query preference. Section 5.3 separately examines a benchmark that includes no intervention. Because the model and surface choices come from the same candidate set, surface-relative regret is bounded by $s(c)$. Realized regret is bounded by $s(c)$ plus reference-surface error at the chosen and environment-best actions. The ranking of realized regret, the tightness of this bound, and gains over frozen shuffled, random-feature, and cross-run controls are empirical questions.

# 3 Capture Without Correct Propagation

We first apply the matched-pulse audit to tasks from the DeepMind Control Suite (Tassa et al., 2018). The severe verdict requires $R^2_{\text{prop}}$ and the upper endpoint of its whole-family 95% interval to be below zero. It is intentionally stronger than failing to match the executed-endpoint oracle.

## 3.1 Primary Cheetah dissociation

Three frozen LeWorldModel checkpoints (Maes et al., 2026) are evaluated on Cheetah task-policy support. The checkpoints come from three separate training-run directories, with unverified seed independence (Appendix B). Each evaluation family supplies three roots and six matched actuator contrasts at a physical horizon of five environment steps. Each contrast holds one named actuator at an exact $\pm 1$ for all five steps, leaving every other action coordinate at its logged baseline; LeWorldModel advances the horizon as one native predictor transition over the five-step block, its trained frame skip, with no intermediate imagined states. The query is endpoint torso speed. Readouts and gate conventions are frozen beforehand, the capture and resolution gates are evaluated on 48 calibration families, and each candidate is then scored on 96 evaluation families.

All three candidates pass root capture and real-effect resolution, whose thresholds require detectable signal. For Candidate 01, the reported zero-referenced $R^2_{\text{real}} = 0.127$ leaves residual squared error at about 87.3% of the zero-effect baseline error; the denominator is uncentered effect energy, not variance about the mean. Propagation comparisons use the same frozen readout and baseline (Appendix A). All three candidates' imagined effects then fail both confirmatory comparisons. Predicted-effect $R^2$ ranges from $-6.33$ to $-3.36$, with every 95% interval below zero, and all paired oracle deficits are positive.

## 3.2 Misalignment and excess magnitude in decoded effects

Across the three checkpoints, the complete predicted feature change stays aligned with the real change: its pooled cosine over all evaluation contrasts and feature coordinates is $0.85$–$0.89$. Passed through the frozen torso-speed coordinate of the task readout, however, the predicted and real scalar effects have uncentered correlations of $-0.13$ to $0.08$ and predicted-to-real RMS ratios of $1.7$–$2.6$ (Figure 2). Both statistics compare scalar responses stacked across evaluation contrasts. With uncentered correlation $c$ and RMS ratio $\gamma$, Equation 4 expands to $R^2_{\text{prop}} = 2\gamma c - \gamma^2$: weak alignment and excess gain jointly explain the negative scores even when the complete feature changes align closely.

## 3.3 Replication and boundaries

The same audit was frozen in advance for three further model groups (Appendix Tables 1 and 3). A prospective Walker Walk test uses three independently seeded LeWorldModel runs and a two-coordinate query of endpoint forward speed and torso uprightness. All three pass joint capture and real-effect resolution, then show below-zero imagined effects with positive oracle deficits: the severe form recurs on a second locomotion morphology.

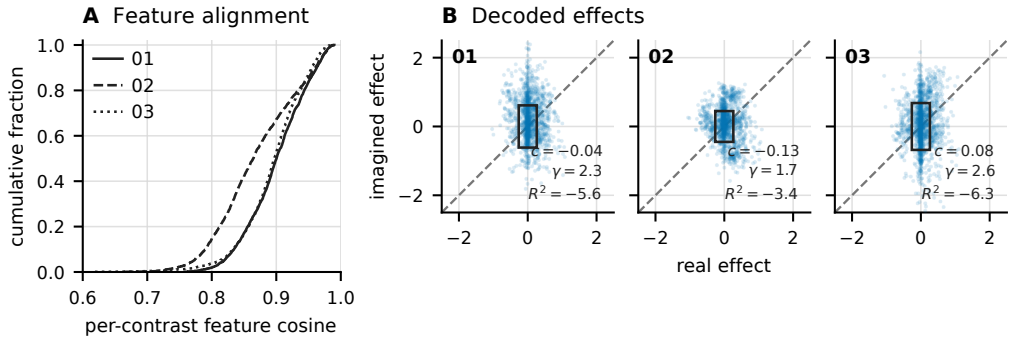



Figure 2: **Aligned feature changes, misaligned decoded effects.** Three LeWorldModel checkpoints on Cheetah, 1,728 evaluation contrasts each. (A) Per-contrast cosine between the predicted and real change in the full predictor state; most contrasts exceed 0.8. (B) The same contrasts passed through the frozen torso-speed readout: decoded imagined effect against the executed effect, in development-scaled units. If propagation were correct, points would lie along the identity (dashed). Instead the imagined effects are 1.7 to 2.6 times larger in RMS (box) and nearly uncorrelated with the real effect; each panel lists the uncentered correlation $c$, RMS ratio $\gamma$, and the resulting $R^2 = 2\gamma c - \gamma^2$. Three points in checkpoint 03 lie beyond the plotted range.

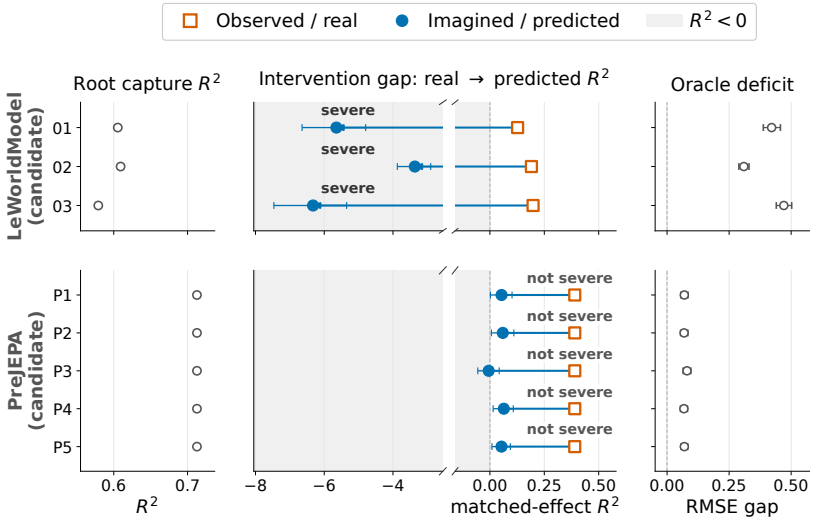



Figure 3: **Cheetah capture and matched-effect propagation.** Rows are model families; intervals resample complete environment families. Left: current-query capture. Middle: real effects (open squares) and imagined effects (filled circles); the broken axis omits no estimate. Right: positive oracle deficit means that the imagined endpoint reads out worse than the executed endpoint. LeWorldModel captures torso speed but distorts and amplifies its matched effect; PreJEPA retains an oracle deficit without the severe below-zero form. Labels index checkpoints or predictors (training units: Appendix B).

Finger Spin broadens the task boundary. Three LeWorldModel seeds, evaluated on endpoint negative hinge velocity, all pass the two prerequisites and have positive oracle deficits. Only one has a severe below-zero verdict; the other two retain positive but oracle-inferior effect signal.

A prospective Cheetah test of five independently initialized action-conditioned PreJEPA predictors over a shared frozen DINOv2 backbone (Oquab et al., 2024) gives the complementary model-family boundary (architecture and training details in Appendix C). The contrast changes both the dynamics-training formulation and the visual backbone (end-to-end learned features against predictor heads over frozen pretrained features). The comparison is therefore categorical (severe versus not severe) and does not isolate either factor. All five predictors pass the prerequisites and all five are worse than the oracle, but none has the severe below-zero form. Their predicted-effect $R^2$ values cluster near zero ($-0.005$ to $0.064$, against real-effect resolution $0.390$); this pattern is consistent with attenuation,

which a separate evaluation on a common set of initial states and actions measures directly for all five predictors (Appendix C). Because a near-zero predicted effect receives a near-zero zero-referenced score, this attenuated response can avoid the severe verdict while still being inaccurate.

Appendix Table 3 retains every candidate estimate.

# 4 Matched-Effect Propagation in TD-MPC2

## 4.1 Cheetah under a matched-shell protocol

A second Cheetah test uses action sequences of fixed norm (action shells). A full-quadratic environment response surface passes its gate on the shells; the fit over the full action ball does not. Separate scalar readouts decode the root query and paired endpoint effects.

The comparison includes three independent four-million-step TD-MPC2 runs (Hansen et al., 2024), three LeWorldModel runs, and five PreJEPA predictors, at two radii and two root supports. The inner radius at policy-generated roots fails the environment gate. Only TD-MPC2 on environment-random roots passes both candidate gates, yielding six passed-gate run-radius evaluations. The effect gate fails at policy-generated roots at the outer radius and for LeWorldModel/PreJEPA on environment-random roots (Appendix D).

All six passed-gate TD-MPC2 candidate-radius evaluations have imagined-effect $R^2$ between $-0.771$ and $-0.267$, with all intervals below zero, and paired oracle deficits of $0.023$–$0.049$, with all intervals above zero (Figure 6). Imagined effects in these passed-gate evaluations are worse than no effect and worse than executed-endpoint decoding. Environment qualification, per-run estimates, geometry, and the broad-readout sensitivity appear in Appendix D.

## 4.2 Matched observed and imagined effects on Finger

For three independently trained Finger models, we apply the same frozen affine readout to observed and five-step imagined endpoints of 64 action directions at each of three roots in 96 families (Figure 4). This descriptive comparison uses the large radius, where the preceding readout test recovered effects accurately; small-effect decoding generalized poorly despite retaining some predictive signal.

Observed-effect centered $R^2$ is $0.932$–$0.952$, versus $-1.168$ to $-0.922$ for imagined effects; the paired RMSE increase is $0.862$–$0.930$, with all intervals positive. After removing each root's mean effect, real scores stay at $0.929$–$0.952$ while imagined scores are $-0.801$ to $-0.581$: the failure includes action-dependent responses, beyond level differences between roots. In the 162 exactly-flat cells among 288, imagined predictions invent variation (mean within-cell SD $0.287$–$0.366$).

Sign and rank statistics require care because exactly zero effects account for $64.47\%$ of observations; sign accuracy restricted to nonzero effects is $0.522$–$0.534$. Half of the 126 nominally rankable cells have true-effect SD below the replay measurement floor. Appendix E retains these diagnostics and the unequal-information surface comparison.

# 5 From Action-Effect Error to Decision Error

We test $s(c)$ on 96 new Cheetah environment families, balanced between radii, with three initial states per family and six candidate actions per state. Predictions for three frozen TD-MPC2 runs were recorded before a single evaluation pass; uncertainty resamples whole families within each run and radius. Pilot results informed the six-action design before development data were examined; the pilot is excluded. We then report separate post-hoc analyses of the Cheetah benchmark and signed-effect selection on Finger.

## 5.1 Held-out regret within the candidate set

The discrepancy score ranks realized regret within the candidate set in all six candidate-radius evaluations (Appendix Table 7). Spearman correlations range from $0.353$ to $0.536$, with every 95% interval above zero. AUROC for the frozen high-regret event ranges from $0.723$ to $0.840$, with every interval above $0.5$. High-regret false acceptance ranges from $3.47\%$ to $9.72\%$.

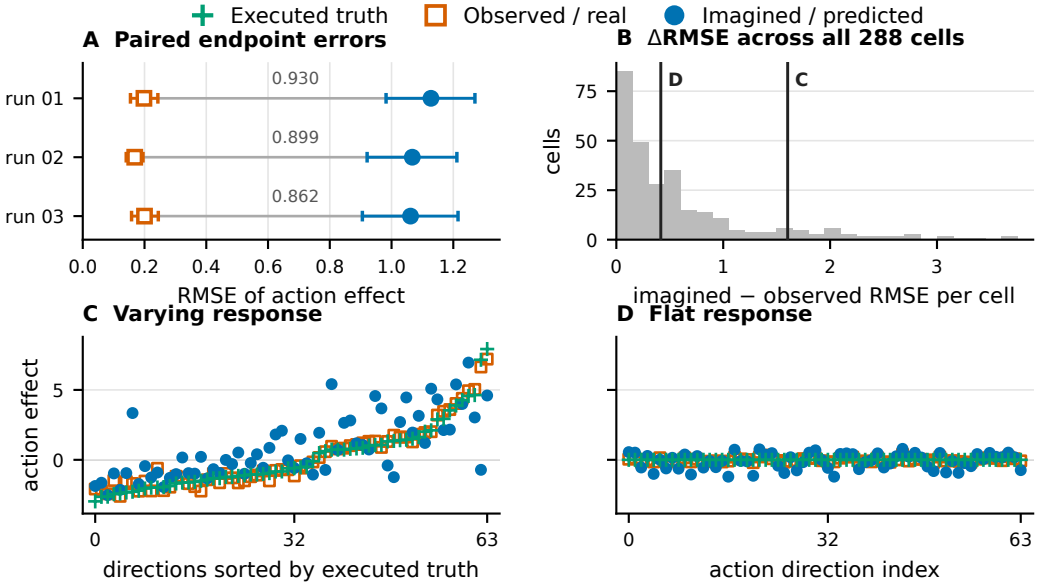



Figure 4: **Finger/TD-MPC2 endpoint effects.** (A) Paired errors by run; labels give the imagined-minus-observed RMSE (intervals in Table 6). (B) Imagined-minus-observed RMSE across all 288 run-01 cells; vertical lines mark the cells shown in C and D. (C–D) Selected varying and flat responses on a shared effect scale; C orders the 64 directions by executed truth. Green: executed truth; orange: observed-endpoint decoding; blue: imagined decoding. The same readout recovers observed effects accurately but fails on imagined endpoints.

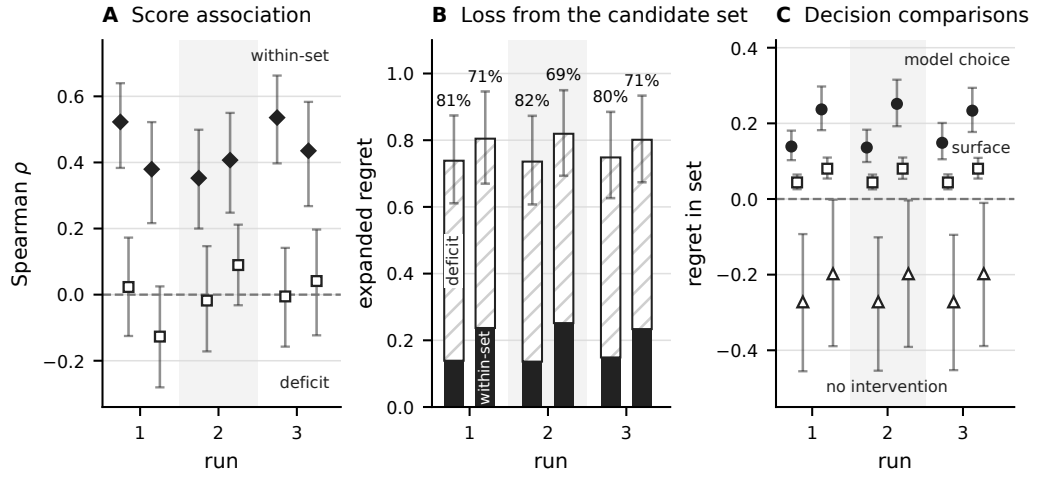



Figure 5: **Choice error, action-set deficit, and decision comparisons.** Within each run, the left value is the inner radius and the right the outer. (A) The score tracks regret within the candidate set; the intervals for its association with the set's deficit relative to no intervention span zero. (B) Bars decompose mean expanded regret; error bars describe the total and labels give the deficit's share of summed regret. (C) Model choices incur more regret than surface choices; negative values for no intervention mean it outperforms the best candidate action on average. A–B are post-hoc; 95% intervals resample whole families within each radius.

Paired Spearman gains over action-shuffled and cross-run controls are positive in all six evaluations. Gains over a 495-dimensional random-feature control are clear in only three of six. That control is capacity- and rank-matched (Appendix F). Point AUROC differences favor the discrepancy score in all six evaluations, but paired AUROC intervals were not part of the frozen analysis.

### 5.2 Decision comparisons

The response-surface form, fitting roles, and thresholds were frozen without using these held-out executed outcomes. The model-selected action has mean query regret $0.136$–$0.252$ across the six candidate-radius evaluations. Direct choices from the qualified reference surface are lower: $0.0439$ at the inner radius and $0.0801$ at the outer radius, with non-overlapping intervals in every candidate-radius comparison (Figure 5C). The analytic bound is loose: mean slack is $2.09$–$2.33$ (Appendix A and Table 8).

Regret intervals overlap in all 48 same-radius comparisons between models that failed a gate and TD-MPC2 runs that passed. A failed gate therefore marks a readout that could not be resolved, not a demonstrated planning deficit. The paired one-million-step readout also fails its gate, so sensitivity to training stage (one-million versus four-million steps) is untested (Appendix F).

### 5.3 When no intervention is the better option

The best candidate action is worse on average than no intervention at both radii ($-0.2721$ inner and $-0.1978$ outer, with both intervals below zero). A post-hoc re-analysis of the stored evaluation outputs measures the loss omitted by the original regret benchmark. Let $r(c)$ be regret within the six-action set and $B(c)$ its best executed effect relative to no intervention. The original choice's regret against the expanded benchmark is

$$R_0(c) = r(c) + D(c), \qquad D(c) = \max\{0, -B(c)\}. \tag{7}$$

The *action-set deficit* $D(c)$ persists even when the best candidate action is chosen. It is shared across runs at a given initial state; the original rule could select only the six candidate actions.

The deficit accounts for 80.1–81.5% of summed expanded regret at the inner radius and 69.3–70.8% at the outer radius (Figure 5B). The score does not track it: Spearman estimates between score and deficit range from $-0.127$ to $0.089$, and every interval spans zero (panel A). Changing the benchmark lowers the correlation point estimate in all six evaluations; paired decreases exclude zero in four, including under the supplementary adjustment for 12 comparisons (Appendix F.1).

### 5.4 Selecting positive effects on Finger

A separate post-hoc analysis uses three frozen Finger runs at radius 0.75, with 192 families and 576 cases per run. An offline rule keeps the model's selected action when $m - |m - p| - b > 0$, where $m$ and $p$ are the model's and a quadratic surface's predicted effects relative to no intervention, and $b$ is an allowance calibrated on development data. Otherwise it selects no intervention. The rule tests only whether the selected action is expected to help; a positive but suboptimal action is kept.

The rule keeps 1.6–4.0% of cases and gives mean development-scaled query effects of 0.075–0.100 relative to no intervention, with all three 95% intervals above zero. These means are over all 576 cases, with zero assigned when the rule selects no intervention. Paired differences from always selecting span zero in every run. For every case the rule declines, the sign of the selected action's effect is undetermined under the calibrated bounds; no selected action, and no whole candidate set, is shown to be negative (Appendix F.2).

## 6 Related Work

**World-model objectives and rollout error.** Latent world models expose encoded states and native open-loop transitions to planners or imagined-rollout learners (Hafner et al., 2019; 2020; 2023; Hansen et al., 2022; 2024). Value-aware objectives target planning-relevant accuracy (Farahmand et al., 2017); value equivalence matches Bellman updates on specified policy and function classes without generally requiring physical metric preservation (Grimm et al., 2020; 2021). Predictive and control objectives can disagree (Lambert et al., 2020), and replanning can mask open-loop drift (Appendix G). Model error can compound under a planner's induced distribution and be exploited where the model is wrong (Ross & Bagnell, 2012; Talvitie, 2017; Kurutach et al., 2018). Our audit complements these objectives by testing whether a model reproduces a declared physical action effect.

**Structural and interventional evaluation.** Structural evaluations show that next-step success need not imply coherent transition structure (Vafa et al., 2024). Recent work also argues for interventional reasoning and rollout validity as central world-model criteria (Yu et al., 2026), probes learned simulator states (Zhang, 2026), and measures action sensitivity in generative interactive models (Bruce et al., 2024; Arai et al., 2024). Our audit adds matched environment endpoints and a strict gate order: query capture and real-effect resolution must pass before imagined propagation is interpreted. The environment-endpoint oracle further separates transition error from the readout's own error.

**Representation and operator diagnostics.** Linear probing motivates controls for capacity and selectivity (Alain & Bengio, 2017; Hewitt & Liang, 2019; Belinkov, 2022). Readability does not imply behavioral use (Elazar et al., 2021; Vakalis, 2026). Koopman and dynamic-mode methods provide language for finite observable families and lifted operators (Koopman, 1931; Williams et al., 2015; Brunton et al., 2022). We use a finite-horizon environment-side response surface fitted to the tested actions.

## 7 Scope and Limitations

The model and task roster appears in Appendix Table 1. Finger TD-MPC2 results use radius 0.75 and five steps; small-radius readout qualification still fails. Models were selected by eligibility rules, so architecture rankings and failure prevalence cannot be inferred. All lack observation reconstruction objectives; an explicit decoder's effect on propagation is untested. The primary Cheetah checkpoints have distinct weights but incomplete records of training independence. Their three-of-three severe result cannot establish seed-invariant severity; independent LeWorldModel Finger seeds show variation within one task.

The results apply to the tested queries, horizons, supports, and native interfaces. Pulse and shell tests use different action designs, environment targets, and readouts, restricting cross-protocol comparisons to the direction of the findings. Differences among tasks also confound morphology, policy, action dimension, support, duration, and training protocol.

Capture depends on the chosen readout family and measures decodability rather than the planner's use of that information. Failed capture or effect resolution leaves propagation and planning ability unknown. The cause of the Finger failure is not identified: inaccurate dynamics, drift into latent states where the readout no longer works, or both could produce it.

The held-out study has no preregistered binary predictive criterion. The score's advantage over the random-feature control is inconsistent across runs and radii, leaving its dynamics-specific contribution unresolved: initial-state geometry or variation in action responses could also contribute. The expanded regret benchmark is a post-hoc decomposition for three frozen runs and describes where the loss sits, not what causes it. The Finger signed rule was likewise chosen after outcomes were known; calibrating it on development data does not change that. Its retained sets are too small to bound harm, and it needs real intervention data to fit its surface. Prospective validation, transfer across model families, and detection of harmful candidate sets are untested, as are planning and return performance.

## 8 Conclusion

Matched interventions expose errors that query decoding and whole-feature alignment miss. On Cheetah, the three primary checkpoints decode the task variable and resolve executed effects, then produce imagined effects that are misaligned and amplified. On Finger, one readout decodes executed effects accurately and imagined effects badly.

On Cheetah, the discrepancy score tracks the model's choice error within a candidate set but not the larger loss from the set itself. On Finger, a conservative signed rule selects a small subset of actions with positive average effect relative to no intervention, with no demonstrable gain over always selecting. The open question these analyses leave is whether an audit can identify harmful selections and harmful candidate sets, not only beneficial ones. Testing that requires prospective evaluation of signed effects against a no-intervention reference.

## AI USE STATEMENT

The research questions, audit design, and gate rules are ours. However we did used AI coding and writing assistants (Claude and Codex agents) for specific parts of this project. The assistants edited text across the manuscript and helped locate and check references. Ultimately, we ran every experiment, verified every number in the paper against the stored outputs, and checked every citation against its source. Any errors are ours.

## REFERENCES


Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations Workshop (ICLR Workshop)*, 2017. URL https://openreview.net/pdf?id=HJ4-rAVtl.

Hidehisa Arai, Keishi Ishihara, Tsubasa Takahashi, and Yu Yamaguchi. ACT-Bench: Towards action controllable world models for autonomous driving. *arXiv preprint arXiv:2412.05337*, 2024. URL https://arxiv.org/abs/2412.05337.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1): 207–219, 2022. doi: 10.1162/coli_a_00422.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, et al. Genie: Generative interactive environments. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4603–4623, 2024.

Steven L. Brunton, Marko Budišić, Eurika Kaiser, and J. Nathan Kutz. Modern Koopman theory for dynamical systems. *SIAM Review*, 64(2):229–340, 2022. doi: 10.1137/21M1401243.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359.

Amir-massoud Farahmand, André M. S. Barreto, and Daniel N. Nikovski. Value-aware loss function for model-based reinforcement learning. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1486–1494, 2017.

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*, volume 33, 2020.

Christopher Grimm, André Barreto, Gregory Farquhar, David Silver, and Satinder Singh. Proper value equivalence. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems 31*, pp. 2450–2462, 2018.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, et al. 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, 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 domains through world models. *arXiv preprint arXiv:2301.04104*, 2023. URL https://arxiv.org/abs/2301.04104.

Nicklas Hansen, Xiaolong Wang, and Hao Su. Temporal difference learning for model predictive control. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8387–8406, 2022.

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

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275.

Bernard O. Koopman. Hamiltonian systems and transformation in Hilbert space. *Proceedings of the National Academy of Sciences*, 17(5):315–318, 1931. doi: 10.1073/pnas.17.5.315.

Thanard Kurutach, Ignasi Clavera, Yan Duan, Aviv Tamar, and Pieter Abbeel. Model-ensemble trust-region policy optimization. In *International Conference on Learning Representations*, 2018.

Nathan Lambert, Brandon Amos, Omry Yadan, and Roberto Calandra. Objective mismatch in model-based reinforcement learning. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pp. 761–770, 2020.

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. URL https://arxiv.org/abs/2603.19312.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Stéphane Ross and J. Andrew Bagnell. Agnostic system identification for model-based reinforcement learning. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

Richard S. Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin*, 2(4):160–163, 1991. doi: 10.1145/122344.122377.

Erik Talvitie. Self-correcting models for model-based reinforcement learning. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 2017.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, et al. DeepMind control suite, 2018. URL https://arxiv.org/abs/1801.00690.

Keyon Vafa, Justin Y. Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model. In *Advances in Neural Information Processing Systems 37*, 2024.

Donna Vakalis. What a world model represents is three questions. arXiv preprint arXiv:2607.06640, 2026. URL https://arxiv.org/abs/2607.06640.

Matthew O. Williams, Ioannis G. Kevrekidis, and Clarence W. Rowley. A data-driven approximation of the Koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25(6): 1307–1346, 2015. doi: 10.1007/s00332-015-9258-5.

Yang Yu, Shiyuan Zhang, Yifei Sheng, Haoxiang Ren, and Haoxin Lin. How should world models be evaluated for embodied decision-making? A decision-making-centric position. *arXiv preprint arXiv:2606.15032*, 2026. URL https://arxiv.org/abs/2606.15032.

Xinyu Zhang. What do world models learn in RL? Probing latent representations in learned environment simulators. In *ICLR 2026, the 2nd Workshop on World Models: Understanding, Modelling and Scaling*, 2026. URL https://arxiv.org/abs/2603.21546.

## A AUDIT FOUNDATIONS

**Why the gates are ordered.** For one matched contrast, insert the executed-endpoint readout between the imagined and literal effects:

$$\|\Delta\widehat{q} - \Delta q\| \le \|\Delta\widehat{q} - \Delta q_{\mathrm{env}}\| + \|\Delta q_{\mathrm{env}} - \Delta q\|. \tag{8}$$

Gate 2 measures the second term, establishing how well the readout resolves executed effects before propagation is interpreted.

Closure coordinates are the environment-side coordinates on which the reference surface is built. For an explicit finite-horizon extension, fix an action sequence and suppose that, in a region containing both trajectories, the closure coordinates satisfy $c_{t+1} = T_t(c_t) + \xi_t$ and $\widehat{c}_{t+1} = \widehat{T}_t(\widehat{c}_t)$. In a fixed weighted norm, assume $\|\xi_t\|_W \le \tau_{C,t}$, $\|\widehat{c}_0 - c_0\|_W \le \epsilon_{\mathrm{cap}}$, $\|\widehat{T}_t(v) - T_t(v)\|_W \le \epsilon_{\mathrm{op},t}$, and $\|T_t(u) - T_t(v)\|_W \le L_t\|u - v\|_W$ throughout that region. Writing $e_t = \|\widehat{c}_t - c_t\|_W$, inserting $T_t(\widehat{c}_t)$ gives the recurrence

$$e_{t+1} \le L_t e_t + \epsilon_{\mathrm{op},t} + \tau_{C,t}. \tag{9}$$

Let $P_{s:h} = \prod_{j=s}^{h-1} L_j$, with $P_{h:h} = 1$. If $y_h = B_h c_h + \rho_h$, $\|\rho_h\|_Y \le \tau_{q,h}$, and $\widehat{y}_h = B_h\widehat{c}_h$, induction yields

$$\|\widehat{y}_h - y_h\|_Y \le \|B_h\|_{W\to Y}\left[\underbrace{P_{0:h}\epsilon_{\mathrm{cap}}}_{\text{root capture}} + \underbrace{\sum_{t=0}^{h-1} P_{t+1:h}\epsilon_{\mathrm{op},t}}_{\text{restricted transition}} + \underbrace{\sum_{t=0}^{h-1} P_{t+1:h}\tau_{C,t}}_{\text{environment approximation}}\right] + \underbrace{\tau_{q,h}}_{\text{query residual}}. \tag{10}$$

For a difference of two endpoint queries, the two arm-specific bounds add by the triangle inequality. The bound depends on the stated coordinate dynamics, approximation errors, and local stability throughout the region. Products of $L_t$ can amplify early errors; we do not estimate these factors or the individual error contributions.

**Held-out regret bound.** For the scalar shell query, let $a_M$, $a_C$, and $a_E$ maximize the model, reference surface, and executed environment displacement, respectively, over the same set of candidate action sequences. Model preference implies $d_M(a_M) \ge d_M(a_C)$, so Equation 6 gives

$$d_C(a_C) - d_C(a_M) \le s(c). \tag{11}$$

Adding and subtracting surface values at $a_E$ and $a_M$, and using $d_C(a_C) \ge d_C(a_E)$, gives

$$d_E(a_E) - d_E(a_M) \le s(c) + |d_E(a_E) - d_C(a_E)| + |d_E(a_M) - d_C(a_M)|. \tag{12}$$

The first inequality follows algebraically when the model and response surface select from the same candidate actions. Whether the second bound is tight and whether $s(c)$ ranks realized regret are empirical questions. The bound's conservatism has two distinct sources. The first is the maximum discrepancy over all action pairs: the maximum in Equation 6 ranges over all fifteen pairs, so a large discrepancy on a pair containing neither the chosen nor the environment-best action still inflates the bound. The second is response-surface prediction error: the two surface-error terms of Equation 12 enter additively and would persist under any envelope. With per-action residuals $z_a = d_M(c,a) - d_C(c,a)$ and $\epsilon_a = d_E(c,a) - d_C(c,a)$, and $x_+ = \max(x, 0)$, the realized slack decomposes exactly into nonnegative terms:

$$\begin{aligned}\big[s(c) + |\epsilon_{a_E}| + |\epsilon_{a_M}|\big] - \big[d_E(a_E) - d_E(a_M)\big] &= \big(s(c) - \big[d_C(a_C) - d_C(a_M)\big]\big) \\ &\quad + \big[d_C(a_C) - d_C(a_E)\big] + 2(-\epsilon_{a_E})_+ + 2(\epsilon_{a_M})_+,\end{aligned} \tag{13}$$

the terms being envelope excess, replacement of the environment maximizer by the surface maximizer, and the two sign penalties of the absolute values. Surface errors that increase realized regret need not contribute slack, so error magnitudes alone are not a slack attribution. The empirical shares of these terms were not computed.

**A tighter bound anchored at the chosen action.** Anchoring the envelope at the model-chosen action, $S_A(c) = \max_{a \in \mathcal{A}_c} |z_{a_M} - z_a|$, also bounds regret: model preference gives $d_C(a_C) - d_C(a_M) = z_{a_M} - z_{a_C} - \big[d_M(a_M) - d_M(a_C)\big] \le |z_{a_M} - z_{a_C}| \le S_A(c)$, and the surface-error terms add as before. Every anchored pair is one of the fifteen, so $S_A \le s$. Because the pair discrepancies are distances among six scalar residuals, $s$ is the residual range and $S_A = s/2 + |z_{a_M} - (u + \ell)/2|$ for extreme residuals $u, \ell$; hence $s/2 \le S_A \le s$, with both ends attained. Anchoring therefore tightens the envelope by at most a factor of two: applied to the reported cell-mean slack fractions of 90–94%, the analytic floors are 80–88%. This is an algebraic comparison; the anchored score was not computed on evaluation data.

### A.1 Synthetic stopping and invariance checks

The ten synthetic cases test stopping logic and invariances. All ten meet their specified checks in the frozen acceptance record. The recipes and thresholds below apply to these synthetic fixtures; empirical gate criteria are given in Section 2.1.

**Common deterministic fixture and scoring.** Use NumPy's seeded `default_rng`, $n = 480$ anchors in 40 balanced, shuffled groups, horizon $H = 3$, and two closure coordinates. Independently draw group centres $\mu_g \sim N(0, 0.7^2 I_2)$, anchor deviations $\eta_i \sim N(0, 0.7^2 I_2)$, and nuisance coordinates $u_i \sim N(0, I_4)$. Set $c_{i0} = \mu_{g(i)} + \eta_i$ and planning state $s_i = (c_{i0}, u_i)Q$, where $Q$ is the orthogonal factor of a Gaussian $6 \times 6$ QR decomposition. Independent equiprobable actions $a_{it} \in \{-1, 1\}$ drive $c_{i,t+1} = c_{it} + a_{it}\gamma$, with $\gamma = (0.45, -0.75)$; no transition or observation noise is added. Capture uses an affine ridge readout with five outer and three inner group-disjoint folds, seed 71, and penalties $\{0, 10^{-6}, 10^{-3}, 0.1\}$. Each coordinate must have centred held-out $R^2 \ge 0.90$ and normalized RMSE $\le 0.20$. All normalization scales are one. At every horizon and coordinate, propagation requires model NRMSE $\le 0.10$, persistence NRMSE $\ge 0.15$, and persistence-minus-model NRMSE $\ge 0.15$. Failed capture leaves propagation unscored even when its raw error is zero.

1. **Exact capture and transition** (seed 19): $\widehat{c}_{it} = c_{it}$; capture and propagation pass.
2. **Wrong operator** (seed 19): initialize at $c_{i0}$, then use $\widehat{c}_{i,t+1} = \widehat{c}_{it} - a_{it}\gamma$. Capture passes and propagation fails.
3. **Absent closure** (seed 23): replace the six pre-rotation state coordinates by independent standard Gaussian noise, but retain the exact rollout arrays. Capture fails and propagation is not scored.
4. **Slow persistence** (seed 29): use $\gamma = (0.02, -0.02)$ and $\widehat{c}_{it} = c_{i0}$. Capture passes; propagation fails both the persistence-resolution floor and the required improvement.
5. **Rank deficiency** (seed 43): draw independent scalar standard Gaussian development/validation samples of sizes 160/120; duplicate each scalar into both columns of each of three horizon blocks. Estimated rank one triggers the declared fallback to full standardized query/pushforward blocks, allowing propagation scoring to continue.
6. **Shuffled action labels** (seed 151): permute complete action sequences across anchors, retaining each target's initial coordinate. Shuffled-action NRMSE must exceed the exact oracle's NRMSE at every horizon and coordinate. The companion controls use seed 991, sample sizes 120/240/480, ranks 1/2, and normalized horizon weights $(0.2, 0.3, 0.5)$.
7. **Stochastic conditional expectation** (seed 181): use 160 anchors, $H = 3$, states/actions in $\{-1, 1\}$, and $\Pr(X_{t+1} = 1 \mid X_t, a_t) = (1 + 0.55X_t + 0.25a_t)/2$. Initial states/actions are equiprobable; the planning state appends three standard Gaussian nuisances to $X_0$. Compare $m_{t+1} = 0.55m_t + 0.25a_t$, $m_0 = X_0$, with means of 4,096 independent sampled trajectories per anchor. Require model NRMSE $\le 0.05$, persistence NRMSE $\ge 0.15$, and improvement $\ge 0.10$; every sampled state must lie in $\{-1, 1\}$.
8. **Off-manifold rollout**: in the same fixture, $m_t + 3$ must fail the finite-state support check (distance tolerance $10^{-12}$). The conditional mean $m_t$ is a valid query expectation but also fails as a deterministic state; sampled states pass with zero tolerance.
9. **Basis rotation and rescaling** (wrong-operator fixture, seed 41): multiply both coordinates and their normalization scales by $(4, 0.25)$, preserving coordinate NRMSE. Separately rotate all coordinates through 0.61 radians, preserving aggregate NRMSE with aggregate scoring.

Compare with tolerance $|x - y| \leq 10^{-8} + 10^{-5}|y|$. The rotation check applies to aggregate scores only.

10. **Split/selection leakage**: require exceptions before scoring for development/evaluation overlap, probe-training/evaluation overlap, evaluation-based rank selection, and validation-based capture-threshold selection. No numerical tolerance permits role overlap.

**Rank selection and exact reproduction.** Standardize using development means and sample standard deviations. Choose the smallest leading rank reaching validation explained variance 0.999. A compressed basis must have its boundary eigenvalue above $0.95$ times the same-position 0.90 quantile of 24 column-shuffled null spectra; retaining the full numerical span needs no null test. Require rank $\geq 2$ and agreement in at least 80% of five 80%-size resamples (seed 811). Failure of these conditions or the variance target triggers the declared fallback. The nondeficient companion uses $z \sim N(0, I_3)$ and blocks $(z_1, z_2)$, $(z_1 + z_3, z_2 - z_3)$, $(2z_1 - z_3, z_2 + z_3)$, with 240/180 development/validation rows, development seeds 3/17/101 and validation seeds offset by 10,000. The executable recipe is `scripts/run_gate_a_acceptance.py`, with fixtures in `src/intervention_operator/diagnostics/synthetic_cases.py` and `stochastic_support.py` in that directory. The frozen settings and ten expected outcomes are `gate_a_config.json` and `acceptance_results.json` under `paper_capture_not_propagation/artifacts/gate_a/` (Python 3.12.13, NumPy 2.3.0). These scripts fix random draw order, fold construction, and rank-estimator details omitted from the compact recipes.

## B Evaluated Models and Comparable Tests

Table 1 identifies the training units and repeated evaluations. LeWorldModel has confirmatory multi-task tests; the TD-MPC2 Finger analyses are descriptive.

Table 1: Evaluated models and supported comparisons. TP is task-policy support; ER is environment-random support.

| Population | Query | Support | Action design | Candidate unit | Comparison |
|---|---|---|---|---|---|
| LeWorldModel Cheetah | torso speed | TP | matched pulse | 3 frozen checkpoints, reported separately | 3/3 severe; 3/3 oracle-inferior |
| LeWorldModel Walker | speed and uprightness | TP | matched pulse | 3 independently seeded runs | 3/3 severe; 3/3 oracle-inferior |
| LeWorldModel Finger | negative hinge velocity | TP | matched pulse | 3 independently seeded runs | 1/3 severe; 3/3 oracle-inferior |
| PreJEPA Cheetah | torso speed | TP | matched pulse | 5 predictor seeds sharing one frozen encoder | 0/5 severe; 5/5 oracle-inferior |
| TD-MPC2 Cheetah | torso speed | ER | matched shell | 3 independently trained runs, each at 2 radii | 6/6 candidate-radius evaluations below zero and oracle-inferior |
| TD-MPC2 Finger | negative hinge velocity | 96 diagnostic families | radius 0.75; action versus no intervention | 3 independent 4M runs | descriptive paired endpoint errors; no qualification gates |
| TD-MPC2 Finger signed rule | negative hinge velocity | 192 evaluation families | radius 0.75; six actions versus no intervention | 3 frozen 4M runs | post-hoc selection of positive effects; separate from endpoint diagnostic |

Table 2 specifies which diagnostic classifications and numerical magnitudes can be compared across tests.

Pulse variants differ in native surface; shells also change action support and the environment target. Gated studies compare qualified effects with zero and executed-endpoint decoding; the Finger diagnostic directly compares matched endpoint errors. All three tasks use action repeat two, but one macro step is $0.02$ s on Cheetah, $0.04$ s on Finger, and $0.05$ s on Walker, so the five-step endpoints occur at $0.10$, $0.20$, and $0.25$ s, respectively.

## C Matched-Pulse Candidate Results and Controls

Table 3 consolidates the matched-pulse results used in Section 3. A severe result requires the imagined-effect $R^2$ point estimate and its upper 95% endpoint to be below zero.

Table 2: Audit protocols and the descriptive endpoint comparison. Physical horizons match executed environment endpoints; one environment step is one action-repeat-two macro step of the wrapped task. Raw scores compare only within a protocol row.

| Protocol | Surface and readout | Fit roles | Native / physical horizon | Supported comparison |
|---|---|---|---|---|
| LeWorldModel pulse | complete predictor context; ridge root and effect readouts | real roots and endpoint contrasts; calibration gates; frozen evaluation | 1 predictor transition / 5 environment steps | gate and verdict counts across its tested tasks |
| PreJEPA pulse | complete patch-grid context; dual ridge readouts | same role-separated pulse design | 1 predictor transition / 5 environment steps | verdict contrast with the Cheetah pulse group |
| TD-MPC2 shell | planning latent; separate scalar root and paired-effect ridge readouts | disjoint development, validation, and evaluation families | 5 native transitions / 5 environment steps | within-shell gate and verdict counts; no raw comparison with pulse scores |
| TD-MPC2 Finger | same affine readout on observed/imagined latent contrasts | readouts fixed before evaluation; 96 diagnostic families | 5 native transitions / 5 environment steps | centered effect scores and paired pooled errors; no new gates |

Table 3: Candidate-level matched-pulse estimates. Intervals use 4,096 whole-family bootstrap resamples over 96 evaluation families per candidate. Oracle values are the oracle deficit of Equation 5; positive is worse.

| Task / model | Candidate | Root $R^2$ | Real $R^2$ | Imagined $R^2$ (95%) | Oracle error (95%) | Severe |
|---|---|---|---|---|---|---|
| Cheetah / LeWM | 01 | .605 | .127 | -5.646 [-6.644,-4.795] | [.388,.457] | yes |
| | 02 | .609 | .191 | -3.364 [-3.873,-2.902] | [.289,.330] | yes |
| | 03 | .579 | .198 | -6.326 [-7.465,-5.349] | [.440,.502] | yes |
| Walker / LeWM | 01 | .713 | .540 | -.351 [-.412,-.292] | [.352,.395] | yes |
| | 02 | .694 | .505 | -.857 [-.985,-.743] | [.476,.539] | yes |
| | 03 | .611 | .530 | -.605 [-.693,-.523] | [.422,.473] | yes |
| Finger / LeWM | 01 | .555 | .763 | -.869 [-1.003,-.760] | [1.368,1.632] | yes |
| | 02 | .539 | .733 | .287 [.212,.352] | [.451,.637] | no |
| | 03 | .554 | .789 | .367 [.338,.391] | [.447,.659] | no |
| Cheetah / PreJEPA | P1 | .713 shared | .390 | .053 [.003,.102] | [.055,.084] | no |
| | P2 | .713 shared | .390 | .059 [.007,.110] | [.054,.084] | no |
| | P3 | .713 shared | .390 | -.005 [-.056,.043] | [.065,.096] | no |
| | P4 | .713 shared | .390 | .064 [.015,.107] | [.054,.082] | no |
| | P5 | .713 shared | .390 | .053 [.009,.094] | [.056,.083] | no |

The three Cheetah checkpoints originate from three separate local training directories with distinct terminal weights, one byte-identical architecture configuration, and one shared dataset path; checkpoint-bound seed and parent-lineage records were not recoverable, leaving training independence unverified. The Cheetah effect readout is two-coordinate, decoding torso speed jointly with a second task coordinate; torso speed is the predeclared primary, and the reported $R^2$ values and the decoded-effect geometry of Section 3.2 use that scalar coordinate alone. Walker and Finger use three independent training seeds. The five PreJEPA predictors are independently initialized but share the same frozen DINOv2 encoder, so their capture estimate is shared.

**PreJEPA candidate specification.** Each PreJEPA candidate pairs the shared frozen encoder, DINOv2-small (ViT-S/14, 22.1M parameters, evaluation mode), with an independently initialized six-layer causal-transformer predictor of width 394 (16 heads, head width 64, MLP width 2,048, dropout 0.1; 19.7M trainable parameters). The predictor consumes three $224\times224$ frames encoded as $3 \times 256$ patch tokens, each token concatenated with a 10-dimensional linear embedding of its 30-dimensional five-step action block (two executed history blocks and one candidate block). Training minimizes mean-squared error over all three shifted frozen-encoder patch-grid predictions, the final shifted target being the true five-step-ahead frame used in evaluation, with the frozen backbone supplying the prediction targets: AdamW at learning rate $5 \times 10^{-4}$, batch 32, ten epochs at constant schedule, full precision after a prespecified mixed-precision finiteness stop, terminal epoch-ten checkpoints only, no early stopping, and no seed replacement. The five runs reach near-identical terminal training losses (0.396–0.402) and all pass a prespecified adequacy gate against the persistence baseline before any matched-intervention contact. The PreJEPA evaluation bank contains 48 probe-training, 48 calibration, and 96 evaluation families; evaluation was opened once, atomically, after all five candidates passed the one-time calibration gate.

In a separate frozen common-bank evaluation, the same five predictors were scored on a shared set of task-policy roots, actions, and endpoints. On the complete surface, the through-origin effect gain $\langle \Delta\widehat{q}, \Delta q\rangle / \|\Delta q\|^2$ was 0.070–0.110, and every 95% whole-family bootstrap upper endpoint was below 0.134, meeting the prespecified attenuation rule (upper endpoint below one); that evaluation's architecture-by-support summaries do not establish a cross-architecture ranking.

**Scalar effect geometry.** Stack the real and predicted scalar effects over evaluation contrasts into nonzero vectors $y$ and $\widehat{y}$. The uncentered correlation is $c = \langle \widehat{y}, y\rangle / (\|\widehat{y}\|\|y\|)$, and the RMS ratio is $\gamma = \|\widehat{y}\|/\|y\|$. Expanding the squared error gives $1 - \|\widehat{y} - y\|^2/\|y\|^2 = 2\gamma c - \gamma^2$. When $c \approx 0$, $R^2 \approx -\gamma^2$. Here alignment concerns vectors of responses across contrasts. Likewise, $1 - R^2_{\text{real}}$ measures residual squared error relative to the zero-effect baseline, whose denominator is uncentered effect energy.

**Role, dependence, and negative controls.** Target construction, probe training, calibration, and evaluation use disjoint environment-family identities. Normalization and gate choices use development data only; candidates are unavailable during target construction. Pulse studies use family-preserving permutations and whole-family bootstrap resampling; their capture and resolution gates require strictly positive $R^2$ with a one-sided 256-draw permutation $p \le 0.05$, and the shell studies' corresponding gates additionally require a strictly positive whole-family bootstrap 2.5th percentile. Their named controls include label and feature shuffles, horizon-zero identity, zero effect, executed-endpoint decoding, wrong action, action null, cyclic actuator, sign swap, readout capacity, and direct feature alignment where defined. Shell studies retain the same role isolation but use their own support qualification, response-surface, action-shuffle, cross-run, and capacity-matched random-feature controls.

## D Matched-Shell Qualification and Candidate Results

The environment-side full-quadratic response surface is candidate-blind and is fitted and qualified on disjoint environment families for the declared fixed-radius shells. The inner and outer action-sequence radii are 0.175 and 0.35; the reference is a 495-term full-quadratic ridge surface fit separately at each radius (ridge $\alpha = 0.001$ inner, 0.01 outer). The study uses 128 readout-development, 64 readout-validation, and 128 shell-evaluation families, split evenly between the two root supports.

Table 4: Environment-side shell qualification. A failed shell stops all candidate propagation claims in that column.

| Root support | Inner radius | Outer radius |
|---|---|---|
| Task-policy | not qualified | qualified |
| Environment-random | qualified | qualified |

The unqualified cells have specific recorded failure modes. Task-policy inner fails exactly one environment gate: family-equal captured variance 0.334 $[0.071, 0.543]$ against the frozen 0.50 threshold, with the effect floor, sign accuracy, split stability, and every control increment passing. Each candidate cell that fails does so at the strictly positive zero-referenced effect-capture gate (the affine readout resolves real effects worse than predicting no effect), often while the within-family permutation null passes, so the limitation is resolution accuracy under the affine readout (Section 2). Support labels specify where environment roots were drawn. Why the task-policy readouts fail their gate was not determined.

The geometry labels in Table 5 follow a pre-specified joint decision rule. After environment-shell and readout qualification, "coherent, wrong" means that the model's self-fit surface qualifies while imagined effects are worse than zero and have a positive oracle deficit; "distorted" means that the self-fit surface does not qualify with the same environmental failure. Run 3 is also distorted at the outer radius, unlike the other two runs.

The self-fit surface is the same 495-feature quadratic fitted to the model's own effects, using measurement actions plus panel A and testing panel B, with ridge penalty 0.001 (inner) or 0.01 (outer). Qualification requires complete, eligible families/roots and action/replay/hash/rank integrity; effect RMS above the environment-random replay floor $3.32 \times 10^{-7}$; family-equal zero-referenced

Table 5: Qualified TD-MPC2 matched-shell estimates. Root and effect columns report gate qualification. Intervals resample complete environment families. Oracle deficit follows Equation 5.

| Run | Radius | Root | Effect | Imagined $R^2$ [95%] | Oracle deficit [95%] | Cosine [95%] | Geometry |
|---|---|---|---|---|---|---|---|
| 1 | inner | pass | pass | -.349 [-.524,-.204] | .0228 [.0194,.0267] | .156 [.120,.191] | distorted |
| 1 | outer | pass | pass | -.267 [-.380,-.171] | .0352 [.0306,.0401] | .189 [.153,.226] | coherent, wrong |
| 2 | inner | pass | pass | -.484 [-.624,-.357] | .0282 [.0240,.0326] | .137 [.098,.177] | distorted |
| 2 | outer | pass | pass | -.539 [-.704,-.396] | .0465 [.0410,.0523] | .182 [.144,.218] | coherent, wrong |
| 3 | inner | pass | pass | -.771 [-1.173,-.468] | .0295 [.0262,.0332] | .227 [.191,.265] | distorted |
| 3 | outer | pass | pass | -.682 [-.948,-.462] | .0489 [.0442,.0541] | .257 [.218,.296] | distorted |

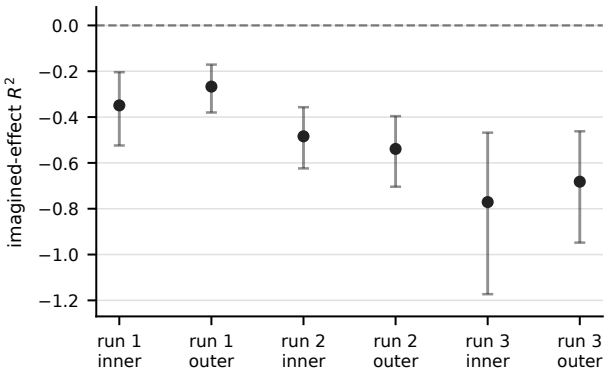



Figure 6: Matched-shell imagined-effect $R^2$ for the six passed-gate TD-MPC2 evaluations, with the values of Table 5. Intervals resample complete environment families; the dashed line is the zero-effect baseline.

$R^2 \geq 0.50$ with 95% lower bound $> 0$; an $R^2$ increment $\geq 0.10$ over each of local-linear, zero-effect, global-linear, query-conditioned-linear, global-secant, family-shuffled, and action-coordinate-shuffled controls; paired sign-flip $p \leq 0.05$ against the local-linear control; split-fit prediction-cosine lower bound $\geq 0.80$; sign accuracy $\geq 0.75$; and RMS magnitude ratio in $[0.50, 2.00]$. Split stability takes the median across roots within family and then a family-bootstrap lower bound for the median across families. Environmental failure here requires an imagined-effect $R^2$ upper bound below zero and a model-minus-oracle RMSE lower bound above zero. The four self-fit-surface failures fail only the increment over the local-linear control: 0.0858/0.0584/0.0401 on the three inner cells and 0.0647 on run 3's outer cell. For these four evaluations, "distorted" describes insufficient improvement over the local-linear control; absolute fit quality is a separate quantity.

A nonlinear-readout sensitivity qualifies no additional candidate. The earlier broad 17-dimensional reconstruction readout, retained only as a named sensitivity, has imagined-effect $R^2 < 0$ in all six qualified evaluations.

## E Matched Finger Endpoint Comparison

The diagnostic uses three independent TD-MPC2 4M runs, 96 families, three roots per family and 64 diagnostic directions at radius 0.75, with five native steps. The query is negative hinge velocity; effects contrast each action with its same-root no intervention. The identical frozen affine readout, including its intercept, decodes observed and imagined endpoint latent differences. Root-capture centered $R^2$ is 0.71–0.80. This comparison reports errors directly, without applying qualification gates.

Unlike Equations 3–4, effect $R^2$ is centered: $1 - \sum_i (\widehat{y}_i - y_i)^2 / \sum_i (y_i - \bar{y})^2$. The paired error difference is $\sqrt{\mathrm{mean}_i[(\widehat{y}_{\mathrm{imag},i} - y_i)^2]} - \sqrt{\mathrm{mean}_i[(\widehat{y}_{\mathrm{real},i} - y_i)^2]}$, using errors pooled across families. Equation 5 instead averages within-family RMSE differences. Each run is separate; 4,096 paired whole-family draws retain all roots and directions. The preceding readout benchmark used different families; the paired comparison uses only the endpoints collected in this diagnostic.

Table 6: Finger endpoint comparison. (a) Centered effect scores and raw errors; brackets are paired 95% family-bootstrap intervals. (b) Post-hoc checks and the secondary, unequal-information quadratic comparison. The surface uses local simulator data unavailable to the model.

*(a) Same readout, observed versus imagined endpoints*

| Run | Real $R^2$ | Imag. $R^2$ | Real RMSE | Imag. RMSE | ΔRMSE [95%] |
|---|---|---|---|---|---|
| 1 | 0.933 | -1.168 | 0.198 | 1.128 | 0.930 [0.801, 1.054] |
| 2 | 0.952 | -0.942 | 0.168 | 1.067 | 0.899 [0.766, 1.033] |
| 3 | 0.932 | -0.922 | 0.200 | 1.062 | 0.862 [0.727, 0.998] |

*(b) Centering, flat responses, sign/rank, and surface comparison*

| Run | Root-centered $R^2$ real / imag. | Flat imag. SD / RMSE | Sign accuracy all / nonzero | Rank Spearman | Surface gap ΔRMSE [95%] |
|---|---|---|---|---|---|
| 1 | 0.929/-0.801 | 0.359/0.551 | 0.1896/0.5336 | 0.1099 | 0.677 [0.577, 0.781] |
| 2 | 0.952/-0.663 | 0.366/0.554 | 0.1865/0.5249 | 0.0911 | 0.617 [0.508, 0.735] |
| 3 | 0.940/-0.581 | 0.287/0.443 | 0.1853/0.5217 | 0.1130 | 0.611 [0.499, 0.729] |

All 162 exactly-flat cells have zero true effects; flat-response SD is averaged equally over cells. Exactly-zero truth comprises 64.47% of rows, so aggregate sign accuracy is zero-dominated. Of 126 nominally rankable cells, 63 have true-effect SD below the frozen replay floor $4.768 \times 10^{-7}$, limiting interpretation of the rank statistics. The per-root quadratic surface is fitted on 160 other simulator directions and has access to local data unavailable to the model. Figure 4 uses run 01 and the first cells in stored order with truth SD above that floor (family/root 0/0) and exactly-zero truth (1/0). This deterministic selection illustrates the two response types; Table 6 summarizes all cases. Source hashes, copied values and selection code accompany `fig14_finger_matched_endpoints.json` in the figure extracts.

## F Held-Out Query-Regret Estimates

The held-out study uses three frozen TD-MPC2 runs, each evaluated separately at the two qualified radii. Each candidate-radius row contains 48 fresh complete environment families and 144 initial-state cases. For each candidate-radius cell, the frozen score threshold $t_s$ is the median of $s(c)$ and the high-regret threshold $t_r$ is the linearly interpolated 75th percentile of realized regret $r(c)$, both over the 72 initial-state cases in 24 action-selection development families at that radius. A high-regret event is $r(c) > t_r$; acceptance requires $s(c) < t_s$, with strict inequalities. AUROC classifies that high-regret event. False acceptance is the joint frequency $144^{-1} \sum_c \mathbf{1}\{s(c) < t_s,\ r(c) > t_r\}$, over all evaluation cases, including rejected ones. The frozen $(t_s, t_r)$ values, in development-scaled query-effect units, are shown below (rounded for display; full precision is retained in `development_thresholds.json`):

| Run | Inner $(t_s, t_r)$ | Outer $(t_s, t_r)$ |
|---|---|---|
| 1 | (0.262162, 0.223484) | (0.413298, 0.269482) |
| 2 | (0.271621, 0.116022) | (0.393930, 0.195200) |
| 3 | (0.341992, 0.121298) | (0.438188, 0.189723) |

The scale is the matched-radius development standard deviation of executed five-step query changes. Both thresholds are fixed from development data before evaluation. Table 7 reports the primary association and high-regret classification estimates.

Table 8 gives the paired Spearman increment beyond the frozen controls and the decision comparators. The action-shuffled control uses the same 495-coordinate full-quadratic design and ridge penalty with a deterministic response derangement; the random-feature control uses 495 deterministic cosine features with the same ridge penalty and fitting rows; the cross-run control is the pointwise median prediction of the other two runs. The frozen analysis specifies paired differences for Spearman correlation only.

As discussed in Section 7, the random-feature result leaves dynamics-specific attribution unresolved. A partial-correlation analysis controlling for root-state norm, initial speed, or the spread of action responses could test these possible confounds; it was not performed.

Table 7: Held-out discrepancy-score performance. Brackets are 95% whole-family bootstrap intervals from 4,096 draws. False acceptance is the fraction of all cases jointly below the frozen score cutoff and above the regret cutoff.

| Run | Radius | Spearman $\rho$ [95%] | AUROC [95%] | False acceptance [95%] |
|---|---|---|---|---|
| 1 | inner | .5226 [.3838,.6428] | .8403 [.7508,.9190] | .0347 [.0069,.0764] |
| 1 | outer | .3795 [.2184,.5151] | .7232 [.6254,.8100] | .0903 [.0417,.1389] |
| 2 | inner | .3526 [.2084,.4930] | .7526 [.6577,.8356] | .0972 [.0486,.1528] |
| 2 | outer | .4071 [.2489,.5545] | .7538 [.6606,.8327] | .0556 [.0208,.0903] |
| 3 | inner | .5356 [.3984,.6556] | .8332 [.7619,.8977] | .0833 [.0347,.1458] |
| 3 | outer | .4353 [.2648,.5825] | .7737 [.6805,.8514] | .0556 [.0208,.0972] |

Table 8: Controls and decision comparisons by run and radius. Panel (a) reports paired Spearman-difference 95% intervals. Panel (b) reports point estimates for model-selected regret, direct reference-surface regret, the comparison with no intervention, analytic bound, and bound slack.

*(a) Paired Spearman increment over control*

| Run | Radius | Action-shuffled | Random-feature | Cross-run |
|---|---|---|---|---|
| 1 | inner | [.1848,.4509] | [.0679,.2759] | [.1326,.4441] |
| 1 | outer | [.1000,.4070] | [-.0750,.1406] | [.1169,.4181] |
| 2 | inner | [.1889,.4524] | [-.0686,.1767] | [.1682,.5214] |
| 2 | outer | [.1316,.3867] | [.0249,.2065] | [.1457,.5470] |
| 3 | inner | [.2534,.4670] | [.0538,.2188] | [.2441,.5487] |
| 3 | outer | [.1867,.4690] | [-.0461,.1247] | [.1568,.4915] |

*(b) Regret within the candidate set and analytic bound*

| Run | Radius | Model | Surface | No intervention | Bound | Slack |
|---|---|---|---|---|---|---|
| 1 | inner | .1388 | .0439 | -.2721 | 2.229 | 2.090 |
| 1 | outer | .2369 | .0801 | -.1978 | 2.495 | 2.258 |
| 2 | inner | .1361 | .0439 | -.2721 | 2.277 | 2.141 |
| 2 | outer | .2516 | .0801 | -.1978 | 2.548 | 2.296 |
| 3 | inner | .1486 | .0439 | -.2721 | 2.285 | 2.137 |
| 3 | outer | .2336 | .0801 | -.1978 | 2.566 | 2.332 |

The six-action design was revised in response to pilot results. The excluded pilot suggested that the number of candidate actions mattered; a subsequent excluded pilot selected six actions before development or evaluation data were examined. A stricter alternative coverage convention would have selected a different design. Tables 7 and 8 report prospective estimates under the amended design (Section 7).

The paired one-million-step affine readout fails qualification in all three runs, and its nonlinear sensitivity cannot replace it; no stage comparison follows. Abstention is also unfavorable: failed-gate LeWorldModel and PreJEPA regret intervals overlap the passed-gate TD-MPC2 range in all 48 same-radius candidate pairings. The negative no-intervention comparison means that continuing without intervention outperforms even the best candidate action on average.

### F.1 Post-hoc regret relative to no intervention

This descriptive re-analysis uses the stored evaluation outputs for the same three frozen TD-MPC2 runs: 48 complete families per radius, with three initial states per family. Its specification and numerical code were frozen before computing the additional summaries, after the original aggregate outcomes were known. Choices, scores, query scales, and eligible cases are unchanged. We resample 4,096 whole-family draws using NumPy PCG64 with seed 2026092102, carrying all roots together and using common draws across runs within a radius. Radii have separate family banks.

Equation 7 compares the original selected action with the best of the six candidate actions and no intervention. In particular, $\max\{B(c), 0\} - [B(c) - r(c)] = r(c) + \max\{0, -B(c)\}$. The deficit is identical across runs for a shared initial state. Its share of expanded regret is $\sum_c D(c) / \sum_c R_0(c)$. The positive part is taken before averaging: the mean deficit is $\mathbb{E}[\max\{0, -B\}]$, whereas negating the mean best effect in Table 8 gives $-\mathbb{E}[B]$.

Mean deficit is 0.600 [0.481, 0.731] at the inner radius and 0.568 [0.444, 0.699] at the outer radius. No intervention beats every candidate action in 61.8% [54.2, 69.4] and 56.25% [48.6, 63.9] of cases,

respectively. These environmental quantities are shared across the three runs. Table 9 gives the resulting loss shares.

Table 9: Post-hoc decomposition of regret against a benchmark including no intervention. Regret values are means in development-scaled query-effect units. Shares are ratios of sums, with 95% whole-family bootstrap intervals.

| Run | Radius | Within set | Set deficit | Expanded | Deficit share (%) [95%] |
|---|---|---|---|---|---|
| 1 | inner | 0.139 | 0.600 | 0.739 | 81.2 [75.5, 86.2] |
| 1 | outer | 0.237 | 0.568 | 0.805 | 70.6 [62.7, 77.1] |
| 2 | inner | 0.136 | 0.600 | 0.736 | 81.5 [75.1, 86.5] |
| 2 | outer | 0.252 | 0.568 | 0.819 | 69.3 [61.0, 76.5] |
| 3 | inner | 0.149 | 0.600 | 0.748 | 80.1 [73.0, 86.1] |
| 3 | outer | 0.234 | 0.568 | 0.801 | 70.8 [63.1, 77.8] |

The original score's within-set Spearman point estimates are unchanged; Figure 5A uses this re-analysis's bootstrap draws, so its intervals differ slightly from Table 7. The deficit association spans zero in all six evaluations. A paired comparison favors within-set over deficit association in all six nominal 95% intervals. The correlation with expanded regret decreases in four evaluations with intervals excluding zero under both nominal and supplementary adjusted coverage (Table 10). The adjustment covers 12 comparisons: six score–deficit associations and six changes from within-set to expanded regret. It uses percentile bounds at per-comparison coverage $1 - 0.05/12$; their extreme tails have limited Monte Carlo resolution with 4,096 draws. No case or undefined bootstrap draw was discarded; all correlations were defined.

Table 10: Post-hoc score associations and paired changes. Brackets are 95% whole-family intervals except in the last column, which adjusts for the 12-comparison family. $\Delta\rho = \rho(s, R_0) - \rho(s, r)$.

| Run | Radius | $\rho(s, D)$ | $\rho(s, R_0)$ | $\Delta\rho$ | Adjusted $\Delta\rho$ interval |
|---|---|---|---|---|---|
| 1 | inner | 0.023 [-0.125, 0.173] | 0.207 [0.065, 0.339] | -0.316 [-0.461, -0.168] | [-0.525, -0.090] |
| 1 | outer | -0.127 [-0.281, 0.025] | 0.152 [-0.028, 0.329] | -0.227 [-0.376, -0.081] | [-0.438, -0.015] |
| 2 | inner | -0.018 [-0.172, 0.147] | 0.114 [-0.039, 0.269] | -0.238 [-0.400, -0.073] | [-0.465, -0.005] |
| 2 | outer | 0.089 [-0.032, 0.212] | 0.365 [0.226, 0.486] | -0.042 [-0.174, 0.094] | [-0.234, 0.148] |
| 3 | inner | -0.005 [-0.157, 0.141] | 0.244 [0.082, 0.403] | -0.291 [-0.436, -0.139] | [-0.503, -0.057] |
| 3 | outer | 0.041 [-0.123, 0.197] | 0.337 [0.148, 0.501] | -0.098 [-0.251, 0.057] | [-0.316, 0.136] |

All original controls are retained. For expanded regret, the paired gain over random features excludes zero only for run 2 at the outer radius (Table 11). The complete summaries for all four scores, all three regret targets, and their paired contrasts are retained in the source archive's `null_decomposition_statistics.tsv`; `fig13_null_decomposition_v1.json` copies the verified summaries used here and in Figure 5A–B.

Table 11: Paired Spearman gains over controls for expanded regret, post-hoc. Brackets are nominal 95% whole-family bootstrap intervals.

| Run | Radius | Action-shuffled | Random-feature | Cross-run |
|---|---|---|---|---|
| 1 | inner | 0.225 [0.082, 0.355] | 0.041 [-0.072, 0.170] | 0.149 [0.010, 0.287] |
| 1 | outer | 0.100 [-0.062, 0.250] | 0.017 [-0.089, 0.122] | 0.096 [-0.087, 0.287] |
| 2 | inner | 0.146 [0.003, 0.289] | 0.004 [-0.121, 0.135] | 0.224 [0.077, 0.383] |
| 2 | outer | 0.191 [0.059, 0.322] | 0.116 [0.006, 0.227] | 0.212 [0.042, 0.391] |
| 3 | inner | 0.144 [0.030, 0.256] | 0.018 [-0.088, 0.122] | 0.203 [0.043, 0.354] |
| 3 | outer | 0.184 [0.022, 0.343] | 0.044 [-0.040, 0.126] | 0.165 [-0.014, 0.336] |

Pairwise action differences in Equation 6 cancel a common shift in the six predicted effects. They measure relative responses within the candidate set, leaving its level relative to no intervention unspecified. An indirect empirical association with the deficit is still possible; the observed intervals span zero. The Cheetah re-analysis changes only the benchmark for the existing choices. The separate Finger rule below evaluates signed effects against a no-intervention reference.

### F.2 POST-HOC SIGNED-EFFECT SELECTION ON FINGER

The signed rule evaluates the negative-hinge-velocity query for three frozen TD-MPC2 Finger runs at radius 0.75 and horizon five. Its 192 evaluation families supply three initial states each. These 576 cases are a separate population from both the Cheetah analysis and the 96-family Finger endpoint diagnostic. No intervention means the five-step zero-action continuation of the environment. Effects use the original development scales. The rule requires real basis-intervention measurements, disjoint from choice actions, to fit its quadratic response surface.

The rule was selected after aggregate evaluation outcomes were known. Its specification and code were frozen before the new calibration, and the calibration was committed before the evaluation stage read the stored evaluation outputs. This separation preserves the development/evaluation roles within the re-analysis; the result is still post-hoc. No new outcomes or model predictions were collected.

For each of the six actions $j$, write the model and surface effects relative to no intervention as $m_j$ and $p_j$, and the executed effect as $y_j$. Retain the model's selected action $a = \arg\max_j m_j$, with the first index breaking ties. Define $\ell_j = m_j - |m_j - p_j|$ and $u_j = m_j + |m_j - p_j|$. On 64 development families, take the maximum over each family's three roots of each residual:

$$\ell_a - y_a, \qquad y_a - u_a, \qquad \max_j y_j - \max_j u_j. \tag{14}$$

For each residual separately, the offset is the 62nd ascending order statistic of the 64 family maxima, clipped below at zero: $\lceil (64+1)0.95 \rceil = 62$. Call these offsets $b_\ell$, $b_u$, and $b_{\mathcal{A}}$. The signed rule keeps the selected action if $L = \ell_a - b_\ell > 0$ and otherwise selects no intervention. Separate upper bounds $U = u_a + b_u$ and $U_{\mathcal{A}} = \max_j u_j + b_{\mathcal{A}}$ report whether the chosen action or the whole candidate set is shown to be negative. This rule assesses the selected action's signed effect independently of whether another candidate action could be better.

Table 12: Signed-rule calibration and retained cases on Finger, post-hoc. The final column counts negative realized effects among retained cases.

| Run | $b_\ell$ | $b_u$ | $b_{\mathcal{A}}$ | Retained | Negative / retained |
|---|---|---|---|---|---|
| 1 | 1.741137 | 1.773530 | 1.773530 | 9/576 | 0/9 |
| 2 | 0.940415 | 0.760510 | 1.567331 | 20/576 | 1/20 |
| 3 | 0.882906 | 0.609014 | 0.682744 | 23/576 | 0/23 |

The nominal calibration level is one-sided 95% and family-marginal under exchangeability of new families with development families. It provides neither a joint two-sided interval nor conditional coverage for selected actions. Complete-family lower-bound violation rates in the evaluation are 1.04%, 5.21%, and 5.73%. The small retained sets in Table 12 are insufficient to establish low harm risk; zero observed negative effects in runs 1 and 3 provide no guarantee about unseen cases.

Table 13 reports means over all 576 cases, assigning zero when the rule selects no intervention. Intervals use the original 4,096 whole-family bootstrap draws, generated with NumPy PCG64 seed 0, retaining all three roots and pairing comparisons within each run. All intervals are nominal 95%. No multiplicity-adjusted claim is made. Undefined conditional quantities and intervals are retained as undefined rather than dropping bootstrap draws.

Table 13: Finger effect relative to no intervention under the signed rule and comparators, post-hoc. Values are mean development-scaled query effects over all cases, with 95% family-bootstrap intervals. The final row is a paired difference.

| Rule / comparison | Run 1 | Run 2 | Run 3 |
|---|---|---|---|
| Always select | 0.1047 [0.0324, 0.1845] | 0.0811 [0.0129, 0.1494] | 0.0939 [0.0211, 0.1696] |
| Model predicts positive | 0.1042 [0.0328, 0.1837] | 0.0835 [0.0158, 0.1520] | 0.0889 [0.0168, 0.1624] |
| Signed rule | 0.0751 [0.0298, 0.1307] | 0.0861 [0.0416, 0.1368] | 0.1004 [0.0522, 0.1566] |
| Signed rule minus always select | -0.0296 [-0.0880, 0.0242] | 0.0050 [-0.0398, 0.0533] | 0.0065 [-0.0428, 0.0619] |

The signed rule's mean effects are positive relative to no intervention in all three runs, but its paired differences from always selecting span zero. Keeping any selected action with $m_a > 0$ retains 90.8–93.2% of cases, and its paired difference from always selecting also spans zero. The rule's

mean difference from always selecting equals the negative loss it avoids minus the positive effects it forgoes. The intervals are consistent with either a net gain or a net loss.

For every declined case the sign is undetermined: neither $U < 0$ nor $U_{\mathcal{A}} < 0$ occurs. The AUROC of $-L$ for a negative chosen effect is 0.618 [0.562, 0.674], 0.545 [0.485, 0.602], and 0.585 [0.531, 0.641]; all whole-set-negative AUROC intervals include 0.5. Strictly negative candidate sets are rare, which limits what the negative-sign test can show: all six effects are negative in only 29/576 cases, although all six are nonpositive in 68.58%. Only 2/576 cases have all six effects below the pre-existing descriptive margin of $-4.77 \times 10^{-6}$. That margin leaves the strict sign rules unchanged. The complete aggregate results and calibration are retained in the source archive.

## G Closed-Loop Control and Candidate Remedies

**Value equivalence and physical queries.** Value equivalence matches Bellman updates for specified policy and function classes; proper value equivalence can retain values sufficient for planning while omitting other environmental detail (Grimm et al., 2020; 2021). Physical metric deformation need not harm those quantities. In particular, an invertible latent coordinate change, with dynamics, reward, value, policy, and readout maps transformed consistently, leaves the represented predictions and decisions unchanged. The audit compares decoded physical effects under a matched readout, allowing arbitrary latent coordinates. Whether such an error affects planning depends on its consequences for relevant rewards and continuation values, consistent with the distinction between prediction and control objectives (Lambert et al., 2020). Value equivalence was not assessed for the tested runs (Section 7).

**Replanning.** A receding-horizon planner executes only the first action of each planned sequence and then replans from a freshly encoded observation of the true environment state, with a learned terminal value guiding the search (Hansen et al., 2022; 2024); repeated re-encoding can mask the multi-step rollout drift measured by the audit.

**Reconstruction as a possible constraint.** DreamerV3 jointly trains an observation decoder, supplying reconstruction gradients to the representation (Hafner et al., 2023). An untested hypothesis is that this constraint reduces loss or drift of task-relevant information relative to a decoder-free objective. Reconstruction alone leaves latent geometry underdetermined: for encoder $E$, decoder $D$, and an invertible, possibly non-isometric coordinate change $A$, $(D \circ A^{-1}) \circ (A \circ E) = D \circ E$. A compensating decoder preserves reconstruction under this change of coordinates. Testing the hypothesis requires a matched reconstructive-model comparison or decoder ablation that evaluates action effects at imagined endpoints.

Untested training directions include penalties that contrast matched action pairs across rollout steps, regularizers aligning model action-derivatives with executed environment responses, and explicit generative observation decoding. The capture-gated audit can evaluate whether these changes improve action-effect propagation.

## H Reproducibility Summary

All primary uncertainty preserves the complete environment family, including all roots and matched actions. Pulse evaluations use 96 families per candidate. The Cheetah held-out study contains 96 eligible families and 288 initial-state cases in total, split into 48 families and 144 cases for each radius; the same cases are repeated across the three frozen runs. Candidate and radius rows are never pooled. Primary intervals use 4,096 whole-family bootstrap draws, while gate tests use the prespecified family-preserving permutation conventions.

The source archive (the supplementary code archive) separates target construction, readout fitting, calibration, prediction, and evaluation roles; evaluation outcomes are opened only after predictions are frozen. Machine-readable final tables retain the full-precision values underlying Tables 3, 5, 6, 7, and 8, and the post-hoc Tables 9–13. Figure-generation scripts record their verified data sources.