VISUAL FIDELITY AND PHYSICAL RELIABILITY IN WORLD-MODEL ROLLOUTS

Anonymous authors
Paper under double-blind review

ABSTRACT

World models are often judged by predicted observations, even when autonomous rollouts are used to reason about physical state. We ask how much a model's native visual prediction error tells us about physical accuracy at the rollout level. Across DreamerV3 and PlaNet, lower visual loss selects rollouts with lower autonomous physical readout error than matched random selection over a broad coverage range, but this ranking does not establish a numerical physical-error scale. We therefore ask what visual loss misses. Around some rollout roots, the queried physical quantity changes strongly under perturbations that produce comparatively little change in the rendered trajectory. A finite-scale query-to-render sensitivity score predicts realized perturbation responses and autonomous physical error beyond visual loss in two independent DreamerV3 cohorts. In a prospectively specified third cohort of 500 episode families, using the same score after visual filtering reduces mean physical error from 0.198 to 0.166 at matched 10% coverage, a 16.0% relative reduction (95% whole-family bootstrap CI for the mean difference, $[-0.0488, -0.0162]$). Finally, we show that positive-loss screening results do not determine whether physical error must vanish as visual loss approaches zero. Visual fidelity is useful for ranking world-model rollouts, but is not sufficient evidence that a particular physical query is reliable.

1 INTRODUCTION

World models are increasingly used not only to predict future observations, but also to support planning and reasoning through imagined futures (Ha and Schmidhuber, 2018; Hafner et al., 2019; 2020; Hansen et al., 2022). Their autonomous rollouts may therefore be queried for position, velocity, or other physical quantities even when the model itself is trained primarily to predict observations. Native visual error is a natural signal to test in this setting: it is already produced by the model's observation interface, summarizes the rendered rollout, and does not require defining a separate physical evaluator before trajectories are compared. If visual error tracked physical accuracy closely, it could provide a simple way to rank or filter rollouts before a more expensive check tied to the physical query. The question is therefore not whether visual and physical error are identical, but how much physical variation is already captured by a metric the world model naturally provides. This leads to a basic evaluation question: **if a rollout has low error in the model's own visual space, how much should we trust the physical quantities read from that rollout?**

Recent work has shown several ways in which visual quality, physical information, and downstream usefulness can diverge. Closed-loop benchmarks evaluate world models through embodied task success rather than visual quality alone, while model-based policy evaluation asks whether imagined rollouts preserve real policy rankings (Zhang et al., 2026; Quevedo et al., 2026). A complementary literature studies which physical variables are recoverable from learned representations and what determines whether they are retained (Tan et al., 2026). Intervention- and consistency-based analyses further distinguish representing the current state from predicting how task-relevant variables change under imagined actions (Vakalis, 2026a; Ruan et al., 2026; Mehta and Baviskar, 2026b). **What remains unclear is how much a model's own visual prediction error tells us about the physical accuracy of an individual rollout, and what information that visual metric misses.**

We study this question using autonomous latent rollouts and fixed physical readouts. In both DreamerV3 and PlaNet, native visual loss contains useful information: retaining the visually best

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.

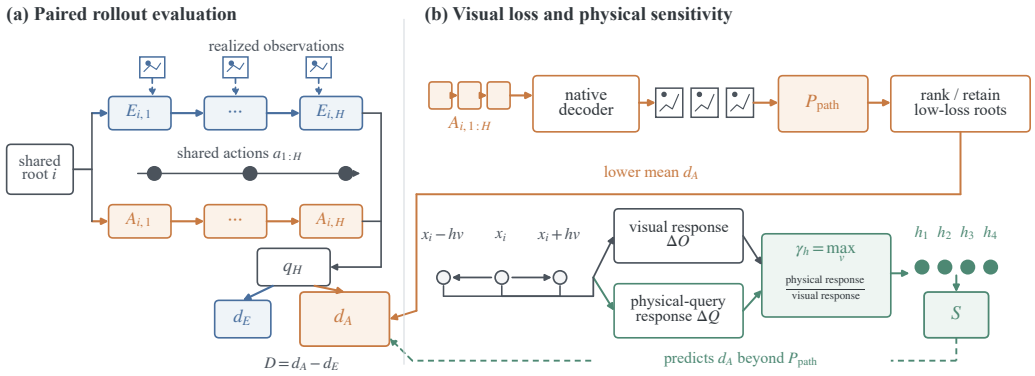

Figure 1: Paired evaluation and the two signals studied here. E and A share the same root, actions, and physical query q_H, giving errors d_E and d_A. Native path loss P^{path} ranks rollouts by visual error, while S summarizes physical sensitivity relative to visual change.

rollouts yields lower autonomous physical error than matched random selection. This matters because the limitation below concerns an otherwise useful metric, not a failed one. The extreme low-loss tail, however, does not establish a numerical physical-error scale.

We trace part of the remaining variation to a simple physical asymmetry: around some states, the queried physical quantity can respond strongly along directions that produce little change in the rendered trajectory. A finite-scale query-to-render sensitivity predicts actual perturbation responses and, beyond visual loss, autonomous physical error in two independent DreamerV3 cohorts. We then ask whether this residual signal changes an evaluation decision rather than merely correlating with error. In a third prospectively specified cohort, a fixed two-stage rule—visual filtering followed by sensitivity refinement—reduces mean autonomous physical error by 16.0% relative to direct visual-only selection at the same final coverage.

The experiments concern only finite, positive visual losses, so they cannot by themselves determine what happens as visual error approaches zero. We show that this limitation is fundamental: two smooth target-error constructions can have the same mean error in the selected low-loss sets above a fixed positive loss floor while differing at the zero-loss limit. **Visual fidelity should therefore be treated as evidence for ranking rollouts, not as sufficient evidence that a particular physical query is reliable.**

2 Evaluation Setup

We compare two representations of the same realized future and evaluate both against the same physical endpoint (Fig. 1). The main experiments then ask how the autonomous physical error varies when rollouts are ranked by the model's own visual prediction error.

2.1 Paired rollouts

Consider a rollout root i, followed by actions $a_{i,1:H}$, realized observations $o_{i,1:H}$, and physical state $q_{i,H}$ at horizon H. Starting from the same root, we form two representations of this future. The observation-aided representation $E_{i,H}$ is updated through the model's native inference interface using the realized observations. The autonomous representation $A_{i,H}$ follows the same actions but receives no future observations.

Both representations are evaluated against the same $q_{i,H}$. They are not information matched: E observes the realized future whereas A must predict it. We therefore use E as an observation-aided reference rather than an equal-information causal control. For DreamerV3 and PlaNet, the two branches correspond to posterior and prior rollouts, respectively. Model-specific representation details are given in Appendix A.

2.2 PHYSICAL READOUT ERROR

Separate readouts with the same architecture are fitted on the observation-aided and autonomous representations. Each readout predicts standardized physical coordinates. For each root, we take the median absolute coordinate error and average this quantity over three independently fitted readouts, giving d_E and d_A. We use d_A as the main measure of autonomous physical error. The branch contrast $D_i = d_{A,i} - d_{E,i}$ is reported when the difference between the two inference interfaces is itself of interest. Full readout, standardization, and fitting definitions are given in Appendix A.

2.3 NATIVE VISUAL LOSS

For DreamerV3 and PlaNet, the native decoder produces an image prediction at every autonomous rollout step. We measure visual error over the full rollout,

$$P^{\mathrm{path}}_{i,m} = \frac{1}{H} \sum_{h=1}^{H} \ell^{(m)}_{\mathrm{vis}}(A_{i,h}, o_{i,h}) \,,$$

where m denotes the world model and $\ell^{(m)}_{\mathrm{vis}}$ is its native squared-RGB loss under the model's image normalization. Because the models have different decoders and normalizations, P^{path} is used only to rank rollouts within model. We omit the model index m when the model is clear from context. Both visual loss and the physical readout are evaluated offline against the realized future, but they require different information: P^{path} uses the model's native decoder and future observations, whereas d_A additionally relies on a declared physical target and a fitted readout. The comparison therefore asks whether the model's own visual metric can stand in for an evaluation tied to the physical query, not whether either quantity is available as an online confidence score.

For a retained fraction α, let $\mathcal{I}_m(\alpha)$ contain the roots with the smallest visual loss. We evaluate their mean autonomous error,

$$\bar{d}^{A}_{m}(\alpha) = \frac{1}{|\mathcal{I}_m(\alpha)|} \sum_{i \in \mathcal{I}_m(\alpha)} d_{A,i},$$

and compare it with selecting the same number of roots at random. Confidence intervals resample source episodes, since several roots can come from the same episode. The native loss is an offline evaluation quantity because it is computed against realized future observations.

2.4 EVALUATION DATA

DreamerV3 (Hafner et al., 2025) provides the main visual-selection, sensitivity, and prospective-selection experiments. PlaNet (Hafner et al., 2019) provides a second RSSM implementation for the visual-selection result using the same Cheetah $qpos + qvel$ query. Additional horizon and readout controls with TD-MPC2 (Hansen et al., 2024) are reported in Appendix C.

The main analyses use $H = 25$ native steps, with action repeat 1 for DreamerV3 and 4 for PlaNet. The DreamerV3 and PlaNet independent evaluation banks each contain 2,000 roots from 100 new source episodes. Multiple roots can originate from the same source episode, so the source episode is the resampling unit for uncertainty. The two sensitivity-prediction cohorts used in Section 4 are disjoint from one another and from the development roots used to define the sensitivity score. A third prospectively specified DreamerV3 cohort contains 1,000 analysis roots from 500 fresh episode families and is used only for the two-stage selection test in Section 4.4. Its rule was fixed before physical-error outcomes were accessed, and it does not recalibrate the sensitivity score. Full data roles and split definitions are given in Appendix A.

A complementary conditional evaluation chooses a model-specific visual-loss operating point on development episodes and applies it unchanged to an independent evaluation bank. Section 3 instead ranks the same bank over a fixed grid of retained fractions. Neither analysis changes the world model or physical readout. The first asks whether one development choice transfers to new episodes; the second asks how well visual loss orders physical error across the bank.

3 What Visual Loss Tells Us About Physical Error

3.1 Low visual loss selects lower-error rollouts

We first ask whether native visual loss contains useful information about autonomous physical error. It does. Across every nontrivial coverage shared by the two models, retaining the lowest-loss rollouts gives lower mean d_A than selecting the same number of roots at random (Fig. 2a). The pattern holds in both DreamerV3 and PlaNet despite their different representations and decoders. At $\alpha = .05$, mean d_A is 0.181 under visual selection versus a random expectation of 0.232 in DreamerV3, and 0.310 versus 0.385 in PlaNet. These values give a representative scale for the effect; the claim is supported by the full common coverage frontier rather than by this single operating point. Because the random comparison retains exactly the same number of roots, the difference isolates how roots are ordered by visual loss rather than a generic benefit from evaluating a smaller subset. Agreement across the different decoders and latent representations reduces dependence on one implementation-specific parameterization. Both systems remain RSSM-family models, so this is replication of the selection phenomenon within that lineage, not evidence of architecture-wide universality.

Native visual loss therefore provides a useful ranking signal for autonomous physical error. Does visual selection identify futures that are easier to read from either representation, or is its benefit larger for autonomous prediction?

3.2 Both branches improve under visual selection

Figure 2b summarizes the branch-specific improvement over random selection. For branch $B \in \{A, E\}$, define

$$G_B = \mathrm{AUC}\big[\bar{d}_B^{\mathrm{random}}\big] - \mathrm{AUC}\big[\bar{d}_B^{\mathrm{visual}}\big]\,,$$

where AUC is trapezoidal integration over the fixed common coverage grid. Positive G_B means that visual selection lowers the integrated branch error. In DreamerV3, $G_A = .0243$ and $G_E = .0034$; in PlaNet, $G_A = .0378$ and $G_E = .0181$.

The observation-aided branch also improves, indicating that visual loss partly ranks futures that are generally easier to read physically. The larger reduction in d_A shows an additional benefit for the autonomous branch, so the smaller E/A difference is not produced by deterioration of the observation-aided reference. We next ask whether this improvement continues into the lowest-loss tail.

3.3 Selection does not establish a physical-error scale

We distinguish useful subset selection from a monotone physical-error scale. Figure 2a asks whether visual selection lowers mean error relative to random selection at the same coverage. Figure 2c asks whether the selected mean continues to decrease as the retained fraction shrinks into the lowest-loss tail. Neither analysis calibrates the numerical value of visual loss to physical error.

In DreamerV3, d_A decreases over the broad part of the curve and then levels off at the smallest retained fractions. PlaNet is more visibly non-monotone in its point estimates: the 2.5% subset has higher mean d_A than the 5% subset. The corresponding marginal intervals overlap, so these point estimates do not establish a population-level reversal.

Visual loss is therefore useful for selecting better rollouts, but its numerical value does not certify their physical accuracy. We next ask what information about physical error is absent from the visual metric. Full coverage curves, results for both branches, and tables for the low-loss tail are reported in Appendix B.

4 What Visual Loss Misses

Visual loss can miss a physical difference when that difference is only weakly expressed in the observations. We therefore ask a local question around each rollout root: **How much can a small state perturbation change the physical query relative to the rendered trajectory?**

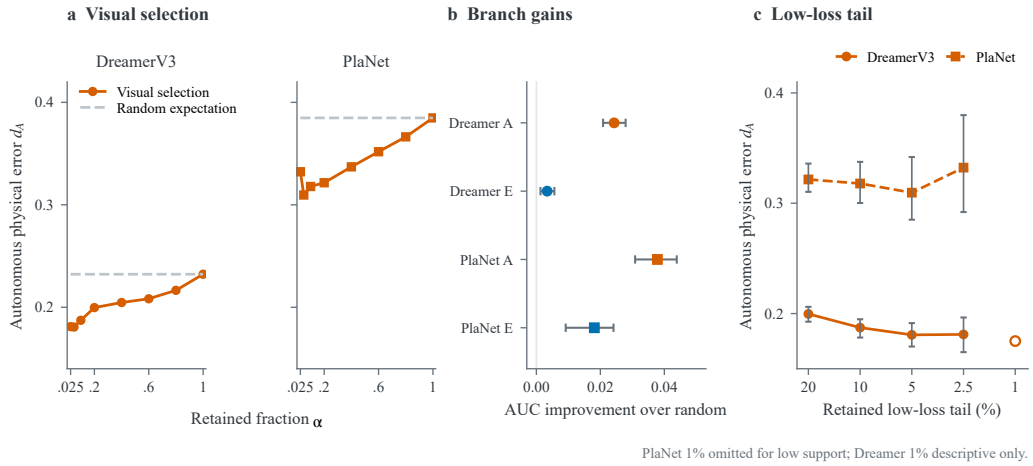

Figure 2: Native visual loss is useful for selection without establishing a numerical physical-error scale. (a) Visual selection lowers d_A relative to same-size random selection. (b) Both branches improve, more strongly for A. (c) Extreme-tail point estimates are non-monotone; DreamerV3 1% is descriptive and PlaNet 1% is omitted for low support.

We study this question in DreamerV3. The object is deliberately finite scale: the rendered trajectory and the queried physical state are nonlinear functions of the environment state, so we use a small set of controlled perturbation sizes and ask which roots are consistently sensitive across them rather than treating one infinitesimal derivative as ground truth.

4.1 Physical sensitivity relative to visual change

At perturbation scale h, we perturb the same root state in opposite directions, replay the same action sequence, and record both the change in the rendered trajectory and the change in the final physical query. These paired responses define two finite-scale quadratic forms: $G^O_{i,h}$ for the rendered observations and $G^Q_{i,h}$ for the standardized query. They summarize how the two outputs respond to the same local directions without requiring the response to be linear beyond the tested scale. We compare the two forms along a common perturbation direction v, using a fixed ridge $\lambda > 0$ in the visual-response denominator:

$$\gamma_{i,h} = \max_{v \neq 0} \frac{v^\top G^Q_{i,h} v}{v^\top \left(G^O_{i,h} + \lambda I \right) v}. \tag{1}$$

The two forms are built from coordinatewise central differences. Along a general direction v, they give quadratic approximations to query and visual responses; Section 4.2 tests these approximations using new perturbations along the maximizing direction. Computationally, Eq. 1 is the leading generalized eigenvalue of the two response forms. The executed finite-difference construction, perturbation coordinates, normalization, and regularization are given in Appendix D.

The absolute magnitude of $\gamma_{i,h}$ changes substantially with perturbation scale, and the maximizing direction is not stable enough to support a scale-invariant local interpretation. We therefore use the calculation only to rank roots at four fixed scales, $h \in \{0.0025, 0.005, 0.01, 0.02\}$. For each scale, F^{dev}_h maps $\log \gamma_{i,h}$ to its percentile in a development set. The multi-scale score is

$$S_i = \text{median}_h \, F^{\text{dev}}_h(\log \gamma_{i,h}) \, . \tag{2}$$

The median is taken over the four scales above. The development mappings are shared by the independent cohorts and are not recalibrated on either one. Equation 2 therefore summarizes relative finite-scale sensitivity without asserting that any single h defines a unique local geometry.

4.2 THE CALCULATION PREDICTS FINITE PERTURBATION RESPONSES

Equation 1 is useful only if its ordering survives an actual perturbation of the environment. We perturb each root in both directions along the selected direction and replay the same action sequence. At each scale, $R_{i,h}^{\text{actual}}$ denotes the realized change in the physical query relative to the realized visual change under this finite intervention; the exact two-sided construction is given in Appendix D. We summarize the rank agreement at each scale by

$$T = \frac{1}{4} \sum_{h} \rho_s \left(\log \gamma_{i,h}, \log R_{i,h}^{\text{actual}} \right). \tag{3}$$

Here ρ_s denotes Spearman rank correlation. The summary in Eq. 3 is positive in both independent cohorts (Fig. 3a). In Cohort 2, $T = 0.410$ with a 95% bootstrap interval over episode families of $[0.300, 0.507]$. Equation 3 tests whether the constructed γ ranks an actually realized query-to-visual response when the simulator is perturbed; stability of the quadratic calculation alone would not establish this. This validates the finite-scale sensitivity construct under the tested intervention. Whether the same development-defined signal also tracks physical error on ordinary, unperturbed autonomous rollouts is a separate question, tested next.

4.3 SENSITIVITY PREDICTS AUTONOMOUS ERROR BEYOND VISUAL LOSS

We now return to the unperturbed rollouts. If S captures physical variation that visual loss misses, roots with larger S should have larger autonomous physical error even after accounting for P^{path}. Both independent cohorts use the same development percentile maps and the same score definition in Eq. 2; neither cohort is used to recalibrate S.

For each cohort, we rank S, visual loss, and d_A, regress the first and third ranks on the visual-loss rank, and correlate the two residuals. The rank formulation avoids assuming a common linear scale for a percentile-based sensitivity score, a decoder loss, and a physical readout error. Uncertainty is computed by resampling episode families and recomputing the ranks and residualizations within each bootstrap replicate, so multiple roots from one source episode do not act as independent evidence. The partial rank association is $0.205\,[0.006, 0.384]$ in Cohort 1 and $0.163\,[0.013, 0.312]$ in Cohort 2 (Fig. 3b). The effects are modest but similar in magnitude across the two independently generated cohorts. In both, larger finite-scale physical sensitivity is associated with larger autonomous physical readout error after linear adjustment for visual-loss rank.

We also condition on three simple measures of trajectory difficulty: root velocity, action magnitude, and realized visual motion. The resulting point estimates retain approximately 91% and 95% of their original magnitude in Cohorts 1 and 2, while both wider bootstrap intervals include zero. We therefore treat this analysis as a descriptive control. Full values are reported in Appendix E.

These analyses show that the fixed score carries information about ordinary rollout error beyond visual loss, but association alone does not show that using S changes which rollouts should be retained. We test that consequence prospectively next.

4.4 SENSITIVITY IMPROVES SELECTION AFTER VISUAL SCREENING

The association above leaves open whether using S improves a selection decision. Before collecting a third cohort, we therefore fixed a two-stage rule: from the full analysis bank, retain the 20% lowest-P^{path} roots, then keep the half with lowest S, giving 10% final coverage. The comparator is direct visual-only selection of the 10% lowest-P^{path} roots. Let

$$\Delta_{\text{sel}} = \bar{d}_A^{\text{two-stage}} - \bar{d}_A^{\text{visual}},$$

so $\Delta_{\text{sel}} < 0$ favors sensitivity refinement.

Cohort 3 contains 1,000 analysis roots from 500 fresh episode families. The two-stage rule gives mean $d_A = 0.1663$, compared with 0.1979 under direct visual-only selection, hence $\Delta_{\text{sel}} = -0.0316$, a 16.0% relative reduction (Fig. 3c). The 95% whole-family bootstrap interval for Δ_{sel} is $[-0.0488, -0.0162]$.

The visual-only comparator already consists of the 100 roots with the lowest P^{path}. The improvement therefore cannot come from selecting a visually lower-loss retained set. This gives the sensitivity signal a prospective decision consequence within the tested DreamerV3 Cheetah setting.

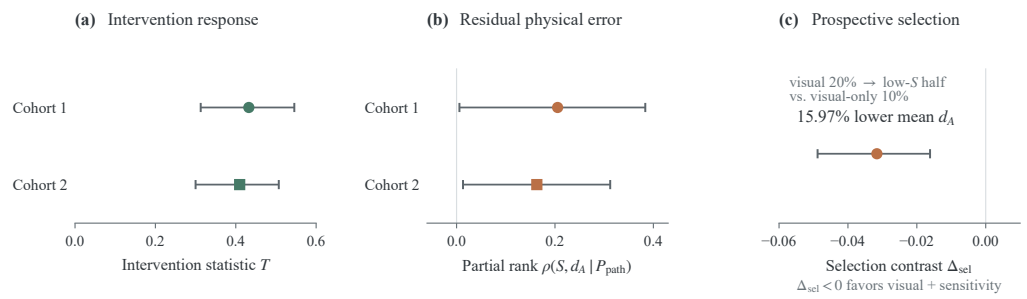

Figure 3: Finite-scale sensitivity links intervention response to rollout selection. (a) Mean Spearman correlation T, averaged over four scales, between calculated sensitivity and realized query-to-visual response. (b) Partial rank association between sensitivity S and autonomous error d_A, adjusting for visual-loss rank. (c) Cohort-3 mean-error difference Δ_{sel} between two-stage and direct visual-only selection at matched 10% coverage. Error bars show 95% episode-family bootstrap intervals.

4.5 Limits of the sensitivity measure

The first limit concerns perturbation scale. In Cohort 2, the rank correlation between $h = .01$ and $h = .02$ falls to $\rho = 0.662$ (Appendix Fig. 7), weaker than the corresponding relationships between neighboring scales. The aggregate four-scale score nevertheless remains stable when any one scale is omitted. This distinction matters because the predictor used in Fig. 3b is the aggregate score in Eq. 2, not the ordering at any single adjacent pair of scales. We therefore interpret S as a summary across the four tested scales, not as evidence that the local geometry is invariant as h changes.

The second limit concerns what error S explains. Its association with the branch difference $D = d_A - d_E$ is unsupported in both independent cohorts. S therefore identifies variation in absolute autonomous physical error that is not captured by visual loss; it does not explain the difference between autonomous and observation-aided inference.

5 Can Finite Low-Loss Data Certify the Zero-Loss Limit?

The experiments above establish two finite-loss facts. First, visual loss is useful for screening; second, a query-specific sensitivity signal can improve selection after visual filtering. Neither result is a numerical certificate. Section 3's extreme tail does not establish that progressively lower visual loss yields progressively lower physical error, and Section 4 remains an average-case decision result over observed positive-loss rollouts. Certification asks a different, worst-case question: whether every sufficiently low-visual-loss rollout must have small error for the queried quantity. A favorable selected mean—even after query-aware refinement—can conceal a small physically inaccurate subset. Because the experiments never observe the zero-loss limit, neither the coverage curve, the sensitivity association, nor the prospective selection gain determines that worst-case behavior. We therefore introduce an estimand that asks what target error remains possible below a visual-loss threshold.

Let Γ be compact and let $V, R : \Gamma \to [0, \infty)$ be continuous visual- and target-error functions, with $\min_{x \in \Gamma} V(x) = 0$. In the empirical setting, these correspond to native visual path loss and autonomous physical readout error. Define

$$\Omega(t) = \sup_{x \in \Gamma:\, V(x) \le t} R(x). \tag{4}$$

A vanishing certificate exists when $\Omega(t) \to 0$ as $t \downarrow 0$, so sufficiently small visual error forces target error to be small as well. For continuous V and R on a compact domain, Eq. 4 vanishes in this limit exactly when $V^{-1}(0) \subseteq R^{-1}(0)$. The condition therefore depends on behavior arbitrarily close to the zero-loss set.

A coverage curve such as Fig. 2, by contrast, records an average over each observed sublevel set. Even knowing the mean error in selected low-loss sets at every positive threshold above some floor leaves open how a small, visually accurate part of the state space behaves near $V = 0$. The observed curve constrains averages over $V \leq t$ but not the pointwise arrangement of target error inside the unobserved neighborhood below the loss floor. Two target-error constructions can therefore preserve the same selected mass and mean error at every observed threshold while differing exactly where certification is decided. The next result makes this separation exact without assuming that the observed screening signal is useless.

Theorem 1. Fix any positive loss floor ε. There exist a compact domain Γ equipped with a probability measure, a continuous visual-error function $V : \Gamma \to [0, \infty)$, and two nonnegative C^1 target-error functions R_{cert} and R_{fail} such that: (i) they agree on $\{V \geq \varepsilon\}$; (ii) for every threshold $t \geq \varepsilon$, they have the same selected mass and the same conditional mean target error on $\{V \leq t\}$; (iii) for every non-saturating threshold $t \in [\varepsilon, \sup_\Gamma V)$, lower-loss selection has smaller mean target error than the full population; yet (iv) R_{cert} admits a vanishing certificate while R_{fail} does not.

Consequently, screening results observed at positive visual loss do not, in general, determine whether a zero-loss guarantee exists without additional structural assumptions. The theorem is an information limit rather than an empirical claim that DreamerV3 or PlaNet retains nonzero physical error at zero visual loss. It shows why increasingly selective finite-loss evaluation cannot by itself turn the relationship observed in Section 3 into a zero-loss guarantee. The explicit construction and proof are given in Appendix F.

6 Related Work

World-model evaluation. World models are increasingly evaluated by what their rollouts support rather than by image quality alone. WorldGym asks whether a learned world model can serve as an environment for robot policy evaluation, while World-In-World evaluates generative world models through closed-loop embodied task success rather than open-loop visual quality (Quevedo et al., 2026; Zhang et al., 2026). WorldModelBench similarly argues that visually plausible generated videos can violate physical constraints that generic video quality metrics miss (Li et al., 2025), while WorldArena evaluates both perceptual quality and functional utility in a unified benchmark (Shang et al., 2026). In a controlled robotics study, Raghavan and Singh (2026) further show that metrics computed from open-loop imagined rollouts can rank state estimators differently from closed-loop control performance. These works establish that visual quality and downstream usefulness should not be conflated. We study a more local version of this problem: whether a model's own visual error can rank the physical accuracy of individual autonomous rollouts.

Physical information in learned world models. Predictive representations have long been studied through the information they preserve for downstream control (Gelada et al., 2019; Schwarzer et al., 2021). More recent work asks directly which physical variables can be recovered from world-model representations and separates what the observations make recoverable from what predictive objectives retain (Tan et al., 2026). PERSIST replaces finite pixel histories with a persistent 3D state, while MIST-WM studies representations designed to retain what a task needs for decision making (Garcin et al., 2026; Feng et al., 2026). Complementary probing studies ask which environment variables are represented and causally used inside learned simulators (Zhang, 2026), and comparisons of reconstruction and semantic representations likewise show that visual reconstruction quality alone does not determine robotic utility (Nilaksh et al., 2026). CALIPER further shows that clean-scene probing can make pretrained and untrained visual representations appear similarly competent at physical inference, and argues that an evaluation's ability to rank representations must itself be demonstrated (Mehta and Baviskar, 2026a). Our experiments do not characterize the full physical content of a representation. They hold the model and physical query fixed and ask how well native visual error predicts the error of a specified physical readout.

Rollout fidelity and interventions. Recent diagnostics evaluate rollout behavior more directly. Operator-on-F and the Intervention Gap compare learned latent transitions with declared observables or matched environment interventions, separating current-state recoverability from the fidelity of imagined dynamics (Vakalis, 2026a;b). Action-state consistency and IMPLY additionally use their diagnostics for candidate selection (Ruan et al., 2026; Mehta and Baviskar, 2026b). Bao (2026) tests a different mechanism by correcting a learned physical invariant during autonomous rollouts.

Rollout-Decoded Reconstruction addresses the related train-test mismatch between observation-anchored decoder training and free-running rollout latents (Shah and Shrestha, 2026). Recent audits of multimodal world models additionally show that visually plausible and explicitly physical outputs from the same model can imply inconsistent physical stories (Zollicoffer et al., 2026). Berger et al. (2026) document diminishing ensemble uncertainty despite growing physical error in RSSM rollouts. Hansen and Wang (2026) use data-coverage-related signals to predict and mitigate world-model hallucination. Here we ask a different question: starting from a model-native visual loss that already ranks rollouts, what query-specific physical variation remains, and can it change selection? We validate query-to-render finite-scale sensitivity through selected-direction environment interventions, test its association with ordinary autonomous physical error after visual-loss-rank adjustment, and prospectively test whether it improves selection after visual filtering.

7 Discussion

Native visual loss remains useful. In both tested world models, it preferentially selects rollouts with lower autonomous physical error, making it a useful first-stage screening signal. Cohort 3 shows why useful is not the same as sufficient: after visual filtering, a fixed query-specific sensitivity score selects a matched-size subset with 16.0% lower mean physical error than direct visual-only selection. The point is not that visual loss fails; it is that the visually best subset can be improved when the second-stage criterion is tied to the physical query. Applications that depend on a particular physical variable should therefore treat visual loss as a screening signal rather than a complete reliability score.

The sensitivity analysis characterizes part of this residual variation. The same amount of visual change can correspond to different amounts of change in the physical quantity being queried. The score S exposes this variation in our DreamerV3 setting, predicts additional autonomous error beyond visual loss, and improves the prospective second-stage selection decision, but it should not be treated as a universal uncertainty measure. Its ordering at each scale is not uniformly stable, and it does not explain the difference between autonomous and observation-aided errors. The result is therefore about physical variation missed by the visual metric, not a complete theory of autonomous rollout error.

The current evidence is bounded by model and access assumptions. Visual selection is tested in DreamerV3 and PlaNet, both based on RSSM-style latent dynamics, while the sensitivity and prospective-selection analyses are established only in DreamerV3 Cheetah and require privileged environment perturbations. Physical error is measured through fitted readouts whose capacity and calibration affect absolute error.

For rollout evaluation, these results support a two-stage design when the downstream application depends on a declared physical variable: use native visual error for inexpensive screening, then apply evidence tied to the query on the retained set. For training and evaluation, the same observation motivates diagnostics or objectives tied to the query whose weighting reflects the consequence of an error for the queried state rather than only its image-space discrepancy. Such signals need not reproduce the particular perturbation calculation used here; they need to be validated against the physical quantity they are meant to protect. The experiments do not identify a universally best objective; they identify information that a useful visual metric can still miss. A practical next step is therefore to develop cheaper signals tied to the query that approximate the diagnostic information exposed here without requiring privileged perturbations. Visual metrics can help decide where to look, but physical claims require evidence about what is being asked of the rollout.

Reproducibility statement

Appendix A specifies the model interfaces, physical readout definitions, data splits, and statistical procedures; Appendices B–E provide the complete numerical analyses, supporting controls, finite-scale construction, independent-cohort prediction, and prospective selection result; Appendix F contains the theoretical statements and proofs; and Appendix G gives implementation and source data details. The supplementary material includes anonymized analysis-level source data and standalone code for reproducing the reported statistics, figures, and tables.

AI USE STATEMENT

Generative AI tools assisted with literature search, code review, plotting code, manuscript drafting and English-language editing, and mathematical proof development and verification. All reported experiments and numerical analyses were run on the authors' research infrastructure; the authors reviewed and verified AI-assisted outputs and take responsibility for the final content.

REFERENCES

Richard Bao. Correcting a learned physical invariant improves world-model rollouts. *arXiv preprint arXiv:2608.23526*, 2026.

Julia Berger, Bernd Frauenknecht, Sebastian Trimpe, and Bastian Leibe. Biased Dreams: Limitations to Epistemic Uncertainty Quantification in Latent Dynamics Models. *Reinforcement Learning Journal*, 7, 2026.

Fan Feng, Yujia Zheng, Minghao Fu, Yongqiang Chen, Guangyi Chen, Kevin Murphy, Biwei Huang, and Kun Zhang. Learning task-sufficient world models by synergizing agentic exploration and structured modeling. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Samuel Garcin, Thomas Walker, Steven McDonagh, Tim Pearce, Hakan Bilen, Tianyu He, Kaixin Wang, and Jiang Bian. Beyond pixel histories: World models with persistent 3D state. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G. Bellemare. DeepMDP: Learning continuous latent space models for representation learning. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2170–2179, 2019.

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*, pages 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 control tasks through world models. *Nature*, 640:647–653, 2025.

Nicklas A. Hansen, Hao Su, and Xiaolong Wang. Temporal difference learning for model predictive control. In *Proceedings of the 39th International Conference on Machine Learning*, pages 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.

Nicklas Hansen and Xiaolong Wang. Hallucination in World Models is Predictable and Preventable. *arXiv preprint arXiv:2606.27326*, 2026.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. WorldModelBench: Judging video generation models as world models. In *Advances in Neural Information Processing Systems*, volume 38, Datasets and Benchmarks Track, 2025.

Aman Mehta and Riya Baviskar. CALIPER: Clean scenes cannot rank physical inference in pretrained visual representations. *arXiv preprint arXiv:2609.08250*, 2026a.

Aman Mehta and Riya Baviskar. IMPLY: Physically anchored consistency for world-model rollouts. *arXiv preprint arXiv:2609.12441*, 2026b.

Nilaksh, Saurav Jha, Artem Zholus, and Sarath Chandar. Reconstruction or semantics? What makes a latent space useful for robotic world models. *arXiv preprint arXiv:2605.06388*, 2026.

Julian Quevedo, Ansh Kumar Sharma, Yixiang Sun, Varad Suryavanshi, Percy Liang, and Sherry Yang. WorldGym: World model as an environment for policy evaluation. In *International Conference on Learning Representations*, 2026.

Dharini Raghavan and Amritpal Singh. Do better imagined rollouts mean better robot control? A controlled study of world-model evaluation under feedback. *arXiv preprint arXiv:2609.02811*, 2026.

Bo-Kai Ruan, Teng-Fang Hsiao, Ling Lo, and Hong-Han Shuai. Is the future compatible? Diagnosing dynamic consistency in world action models. *arXiv preprint arXiv:2605.07514*, 2026.

Yu Shang, Zhuohang Li, Yiding Ma, Weikang Su, Xin Jin, Ziyou Wang, Lei Jin, Xin Zhang, Yinzhou Tang, Haisheng Su, Chen Gao, Wei Wu, Xihui Liu, Dhruv Shah, Zhaoxiang Zhang, Zhibo Chen, Jun Zhu, Yonghong Tian, Tat-Seng Chua, Xin Wang, Wenwu Zhu, and Yong Li. WorldArena: A unified benchmark for evaluating perception and functional utility of embodied world models. *AI Open*, 7:208–226, 2026.

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

Rishi Shah and Rishav Shrestha. Rollout-decoded reconstruction for long-horizon prediction in latent world models. *arXiv preprint arXiv:2608.25017*, 2026.

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.

Donna Vakalis. The intervention gap in latent world models. *arXiv preprint arXiv:2608.29998*, 2026a.

Donna Vakalis. Operator-on-F complements value-equivalence: A planning-time diagnostic for latent world models. *arXiv preprint arXiv:2607.04464*, 2026b.

Xinyu Zhang. What do world models learn in RL? Probing latent representations in learned environment simulators. *arXiv preprint arXiv:2603.21546*, 2026.

Jiahan Zhang, Muqing Jiang, Nanru Dai, Taiming Lu, Arda Uzunoglu, Shunchi Zhang, Yana Wei, Jiahao Wang, Vishal M. Patel, Paul Pu Liang, Daniel Khashabi, Cheng Peng, Rama Chellappa, Tianmin Shu, Alan Yuille, Yilun Du, and Jieneng Chen. World-In-World: World models in a closed-loop world. In *International Conference on Learning Representations*, 2026.

Geigh Zollicoffer, Minh Vu, Rajiv Ranasinghe, and Manish Bhattarai. One model, two physical stories: Auditing misalignment in multi-modal world modeling. *arXiv preprint arXiv:2609.14833*, 2026.

A Experimental Setup and Statistical Definitions

The main text introduces only the quantities required to follow the scientific argument. This appendix gives the model interfaces, physical readout definitions, data splits, and inferential procedures used throughout the experiments.

A.1 World-model interfaces

We evaluate three learned world-model systems. DreamerV3 (Hafner et al., 2025) provides the main visual-selection experiments and the finite-scale sensitivity analyses. PlaNet (Hafner et al., 2019) provides a second RSSM implementation for the visual-selection result. TD-MPC2 (Hansen et al., 2024) is used only for the additional horizon and readout controls in Appendix C.

For each evaluated root i, the observation-aided branch E and autonomous branch A start from the same environment state and receive the same action sequence. E is updated through the model's native inference interface using realized future observations, whereas A advances without receiving those observations. Both branches are evaluated against the same physical endpoint.

For DreamerV3, the representation used by the physical readout concatenates the deterministic recurrent state with the sampled stochastic state. PlaNet uses its recurrent belief together with its stochastic state. TD-MPC2 uses its frozen latent representation. Native action repeats are 1, 4, and 2 for DreamerV3, PlaNet, and TD-MPC2, respectively.

The main DreamerV3 and PlaNet analyses use Cheetah Run and the complete simulator $qpos + qvel$ query available to the executed evaluation pipeline. Their visual-loss and physical-error scales are model specific; no numerical comparison of absolute losses across the two models is intended. Table 1 summarizes the three system interfaces and their roles in the evidence.

Table 1: Experimental systems and roles.

System	Main role	Readout representation	Query	Repeat
DreamerV3	selection; sensitivity	det. + stochastic	full $qpos + qvel$	1
PlaNet	selection replication	belief + stochastic	full $qpos + qvel$	4
TD-MPC2	horizon/readout controls	frozen latent	10-coordinate query	2

A.2 Physical targets and readouts

For physical coordinate j, we standardize the target using statistics from the corresponding fitting data. For branch $B \in \{E, A\}$ and readout fit s,

$$\begin{aligned}
\widetilde{q}_{i,H,j} &= \frac{q_{i,H,j} - \mu_j}{\sigma_j}, \\
e^{(s)}_{B,i,H,j} &= \widetilde{\widehat{q}}^{(s)}_{B,i,H,j} - \widetilde{q}_{i,H,j}, \\
d^{(s)}_{B,i} &= \mathrm{median}_j \left| e^{(s)}_{B,i,25,j} \right|, \\
d_{B,i} &= \frac{1}{3} \sum_{s=1}^{3} d^{(s)}_{B,i}.
\end{aligned}$$

The same target normalization is used for E and A within an experimental setting. The autonomous and observation-aided errors are $d_{A,i}$ and $d_{E,i}$, with branch difference $D_i = d_{A,i} - d_{E,i}$.

Most conditional analyses in the main text use d_A directly. D is used only when the contrast between the two inference interfaces is relevant.

The main DreamerV3 and PlaNet nonlinear readouts are two-layer GELU MLPs of width 1024, fitted separately on the E and A feature distributions. This support-specific fitting prevents the E/A difference from being determined merely by applying a readout trained on one representation

to the other. Table 2 lists the corresponding readout configurations; Appendix C reports explicit support-matching and capacity controls.

Table 2: Physical readout configurations.

Setting	Features	Readout	Role	Fits
DreamerV3 main	det. + stoch.	2-layer GELU, 1024	path/sensitivity	3
PlaNet main	belief + stoch.	2-layer GELU, 1024	80 fit + 20 val fam.	3
TD-MPC2 depth	latent	MLP, 1024	depth	3
TD-MPC2 width	latent	256/512/1024	capacity	3
TD-MPC2 calibration	latent	nonlinear MLP	data sweep	3

A.3 Native visual path loss

For world model m,

$$P_{i,m}^{\mathrm{path}} = \frac{1}{H} \sum_{h=1}^{H} \ell_{\mathrm{vis}}^{(m)} \left(A_{i,h}, o_{i,h} \right).$$

The loss is model native and therefore not compared numerically across architectures. For a retained fraction α, $\mathcal{I}_m(\alpha)$ contains the roots with the lowest path loss, and

$$\bar{d}_m^A(\alpha) = \frac{1}{|\mathcal{I}_m(\alpha)|} \sum_{i \in \mathcal{I}_m(\alpha)} d_{A,i}.$$

Analogous quantities $\bar{d}_m^E(\alpha)$ and $\bar{d}_m^D(\alpha)$ replace d_A with d_E or D.

A.4 Evaluation banks and inferential units

The DreamerV3 development and independent visual-selection banks each contain 2,000 rollout roots from 100 source episodes. PlaNet uses pairwise-disjoint calibration, development, and independent sets with the same family/root counts. The sensitivity development set contains 64 roots from 32 families. Cohort 1 contains 96 roots from 67 families, and Cohort 2 contains 160 roots from 117 families. Fresh Cohort 3 contains 1,000 analysis roots from 500 episode families and is used only for the prospective decision test in Section 4.4.

Several roots can originate from the same source episode. We therefore treat the source episode, rather than an individual root, as the resampling unit and refer to it as an episode family. Table 3 summarizes the role and support of each evaluation bank.

Table 3: Data roles in the main experiments.

Data set	Model	Roots	Families	Role
development screening	DreamerV3	2000	100	visual-loss operating points
independent screening	DreamerV3	2000	100	Section 3
development screening	PlaNet	2000	100	selection calibration
independent screening	PlaNet	2000	100	Section 3
sensitivity development	DreamerV3	64	32	scale/percentile maps
Cohort 1	DreamerV3	96	67	first prediction cohort
Cohort 2	DreamerV3	160	117	independent cohort
Cohort 3	DreamerV3	1000	500	prospective two-stage selection

A.5 BOOTSTRAP INFERENCE

Unless stated otherwise, confidence intervals are percentile intervals from 10,000 episode-family bootstrap replicates. Complete source families are sampled with replacement, and all roots belonging to a sampled family are retained. Quantities that depend on within-cohort ranks, including the partial-Spearman analyses in Appendix E, are recomputed within each replicate rather than reusing ranks from the original sample.

The low-loss-tail analysis additionally reports simultaneous familywise intervals over the supported retained fractions for D. These intervals are used only for the finite-tail statement involving the E/A contrast.

B FULL RESULTS FOR VISUAL SELECTION

This appendix provides the full curves underlying Section 3. The main text emphasizes three conclusions: native visual loss improves selection relative to random choice, both readout branches improve under that selection, and the extreme-tail point estimates do not establish continued monotone improvement in autonomous physical error.

B.1 AUTONOMOUS PHYSICAL ERROR ACROSS COVERAGE

Figure 4 reports the branch-specific coverage curves. Table 4 gives the corresponding numerical values and intervals. Across every nontrivial common coverage in both models, visual selection gives lower mean autonomous error than same-size random selection. The result is not restricted to a single operating point; it persists throughout the common grid from $\alpha = .025$ to $\alpha = .8$. At $\alpha = .05$, DreamerV3 has mean autonomous error 0.1807 under visual selection compared with a full-bank random expectation of approximately 0.2323. PlaNet has 0.3096 under visual selection compared with approximately 0.3848 under random selection. These values illustrate the scale of the selection effect; the claim is based on the full frontier.

Table 4: Autonomous physical error under visual and same-size random selection.

Model	α	n	Visual d_A	95% CI	Random d_A	95% CI	Gain	95% CI
DreamerV3	.025	50	0.1811	[0.1650,0.1965]	0.2323	[0.2271,0.2373]	0.0512	[0.0359,0.0676]
DreamerV3	.05	100	0.1807	[0.1701,0.1913]	0.2323	[0.2271,0.2373]	0.0515	[0.0409,0.0625]
DreamerV3	.10	200	0.1873	[0.1783,0.1949]	0.2323	[0.2271,0.2373]	0.0450	[0.0371,0.0549]
DreamerV3	.20	400	0.1997	[0.1926,0.2061]	0.2323	[0.2271,0.2373]	0.0326	[0.0263,0.0396]
DreamerV3	.40	800	0.2046	[0.1993,0.2100]	0.2323	[0.2271,0.2373]	0.0276	[0.0227,0.0326]
DreamerV3	.60	1200	0.2083	[0.2037,0.2130]	0.2323	[0.2271,0.2373]	0.0240	[0.0199,0.0280]
DreamerV3	.80	1600	0.2165	[0.2124,0.2213]	0.2323	[0.2271,0.2373]	0.0157	[0.0124,0.0186]
DreamerV3	1.00	2000	0.2323	[0.2271,0.2373]	0.2323	[0.2271,0.2373]	0.0000	–
PlaNet	.025	50	0.3323	[0.2920,0.3798]	0.3848	[0.3790,0.3908]	0.0526	[0.0051,0.0930]
PlaNet	.05	100	0.3096	[0.2851,0.3419]	0.3848	[0.3790,0.3908]	0.0752	[0.0430,0.0995]
PlaNet	.10	200	0.3179	[0.3002,0.3375]	0.3848	[0.3790,0.3908]	0.0669	[0.0475,0.0847]
PlaNet	.20	400	0.3216	[0.3103,0.3359]	0.3848	[0.3790,0.3908]	0.0633	[0.0496,0.0741]
PlaNet	.40	800	0.3371	[0.3284,0.3458]	0.3848	[0.3790,0.3908]	0.0477	[0.0399,0.0557]
PlaNet	.60	1200	0.3518	[0.3446,0.3589]	0.3848	[0.3790,0.3908]	0.0330	[0.0284,0.0376]
PlaNet	.80	1600	0.3663	[0.3603,0.3726]	0.3848	[0.3790,0.3908]	0.0185	[0.0151,0.0216]
PlaNet	1.00	2000	0.3848	[0.3790,0.3908]	0.3848	[0.3790,0.3908]	0.0000	–

B.2 BRANCH-SPECIFIC EFFECTS

Using trapezoidal integration on the fixed coverage grid, define

$$\mathrm{AUC}_B^{\pi} = \int_{\alpha_{\min}}^{1} \bar{d}_{\pi}^{B}(\alpha)\, d\alpha, \qquad B \in \{A, E, D\},$$
$$G_B = \mathrm{AUC}_B^{\mathrm{random}} - \mathrm{AUC}_B^{\mathrm{visual}}.$$

Positive G_B means that visual selection lowers the corresponding error.

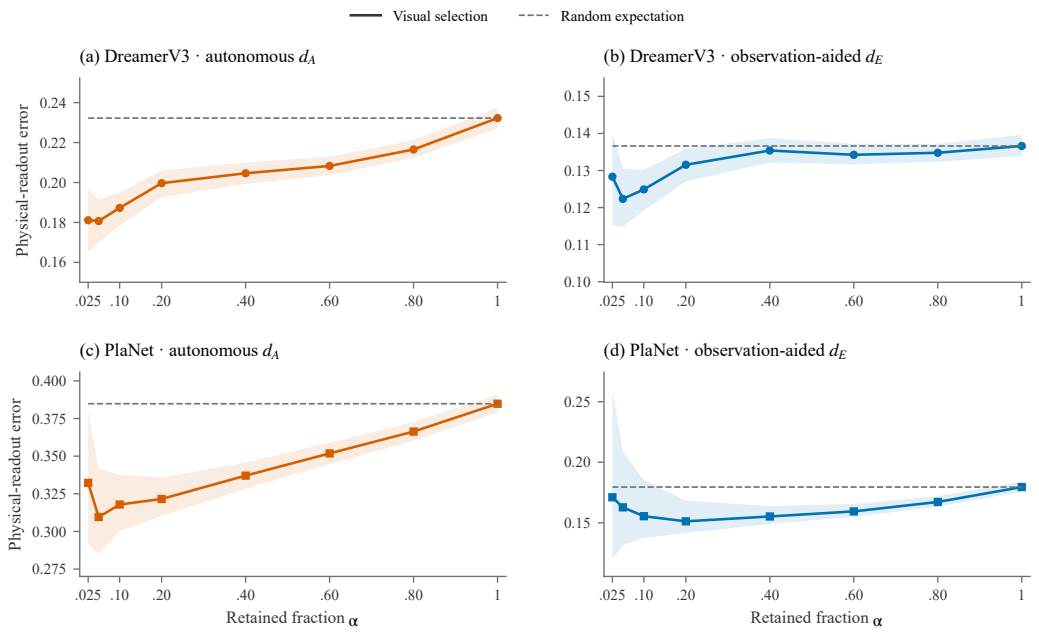

Figure 4: Branch-specific physical error across the visual-selection coverage frontier. Each panel compares selection by native visual loss with same-size random selection. Color identifies the branch (A orange, E blue); line style identifies visual versus random selection. Visual selection lowers both observation-aided and autonomous error over the measured frontier, with the larger integrated reduction on the autonomous branch. Numerical scales are interpreted within model.

For DreamerV3, the gains are $G_A = 0.02430$, $G_E = 0.00337$, and $G_D = 0.02093$; for PlaNet they are 0.03783, 0.01812, and 0.01971, respectively. Thus visual selection improves both branches, with a larger integrated reduction in the autonomous branch. Table 5 reports the gains and bootstrap intervals over episode families.

Table 5: Integrated branch gains from visual selection relative to random selection.

Model	G_A (95% CI)	G_E (95% CI)	G_D (95% CI)
DreamerV3	0.02430 [0.02086, 0.02790]	0.00337 [0.00129, 0.00563]	0.02093 [0.01795, 0.02390]
PlaNet	0.03783 [0.03087, 0.04388]	0.01812 [0.00917, 0.02411]	0.01971 [0.01379, 0.02556]

B.3 Low-loss tails in autonomous error

The strict-tail rows are taken from the same branch losses as the common coverage curves. For DreamerV3, mean autonomous error is approximately 0.1997 at 20% coverage, 0.1873 at 10%, 0.1807 at 5%, 0.1811 at 2.5%, and 0.1751 at 1%. For PlaNet, the corresponding values are approximately 0.3216, 0.3179, 0.3096, 0.3323, and 0.3514.

The supported confidence intervals overlap across the smallest fractions, so the tail fluctuations are not evidence that reducing visual loss causes physical error to increase. They show instead that the point estimates no longer provide a monotone physical-error scale. The 1% rows are descriptive point summaries only; no confidence interval was generated for them. Table 6 reports the full strict-tail values.

Table 6: Autonomous physical error in increasingly selective low-visual-loss subsets.

Model	Retained	n	Visual d_A	95% CI	Random	Gain
DreamerV3	20%	400	0.1997	[0.1926,0.2061]	0.2323	0.0326
DreamerV3	10%	200	0.1873	[0.1783,0.1949]	0.2323	0.0450
DreamerV3	5%	100	0.1807	[0.1701,0.1913]	0.2323	0.0515
DreamerV3	2.5%	50	0.1811	[0.1650,0.1965]	0.2323	0.0512
DreamerV3	1%	20	0.1751	–	0.2323	0.0572
PlaNet	20%	400	0.3216	[0.3103,0.3359]	0.3848	0.0633
PlaNet	10%	200	0.3179	[0.3002,0.3375]	0.3848	0.0669
PlaNet	5%	100	0.3096	[0.2851,0.3419]	0.3848	0.0752
PlaNet	2.5%	50	0.3323	[0.2920,0.3798]	0.3848	0.0526
PlaNet	1%	20	0.3514	–	0.3848	0.0334

B.4 Finite-tail E/A difference

The same low-loss-tail analysis can be expressed through $D_i = d_{A,i} - d_{E,i}$. For DreamerV3, simultaneous familywise 95% intervals remain above zero at every supported retained fraction:

α	$\overline{D}$	95% simultaneous interval
20%	0.06767	$[0.06060, 0.07473]$
10%	0.06184	$[0.05103, 0.07266]$
5%	0.05745	$[0.04502, 0.06988]$
2.5%	0.05350	$[0.03555, 0.07145]$
1%	0.05452	$[0.03481, 0.07423]$

For PlaNet, the supported simultaneous intervals are also positive:

α	$\overline{D}$	95% simultaneous interval
20%	0.16962	$[0.15550, 0.18373]$
10%	0.16158	$[0.14112, 0.18204]$
5%	0.14595	$[0.11777, 0.17413]$
2.5%	0.16245	$[0.12142, 0.20348]$

The PlaNet 1% cell has lower support and is descriptive only. These results concern a finite grid of positive visual-loss thresholds and are not evidence for the $P^{\text{path}} \to 0$ limit.

B.5 Development-selected operating point

The original conditional experiment chose a model-specific path-loss threshold on development data and applied it unchanged to new episodes. In DreamerV3, the selected independent set contained 417 of 2,000 roots spanning 99 of 100 source families. PlaNet selected 379 of 2,000 roots spanning 95 families.

The corresponding mean E/A differences were $0.0677\,[0.0620, 0.0732]$ for DreamerV3 and $0.1696\,[0.1574, 0.1818]$ for PlaNet. These operating points motivated the broader coverage analysis in the main paper but are not treated as the sole evidence for the relationship between visual and physical error.

C Additional Readout and Rollout Controls

The main paper is organized around visual selection and physical sensitivity. The experiments in this appendix characterize how the earlier observation-aided/autonomous contrast depends on horizon, task, physical query, readout fitting, and representation processing. They provide context for the measurement but are not used to define the main selection or sensitivity claims.

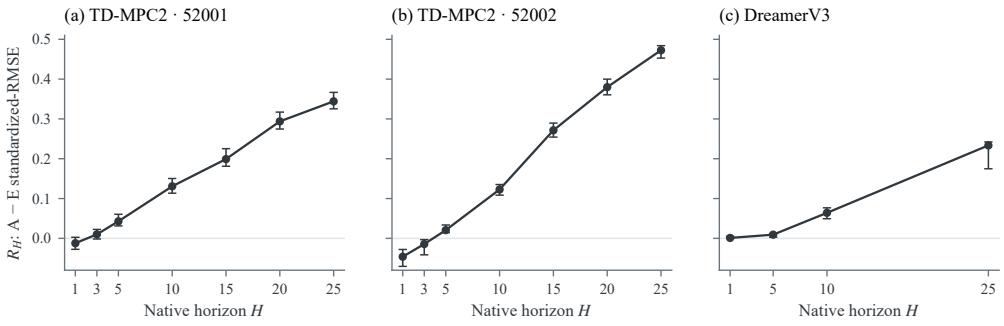

Figure 5: Horizon dependence of the observation-aided/autonomous physical readout difference. R_H is autonomous minus observation-aided standardized RMSE. Error bars are the existing episode-family bootstrap intervals. Horizons, action repeats, and physical queries differ across systems, so only within-system depth trends are interpreted.

C.1 Horizon dependence

For the horizon-level analysis, define

$$\mathrm{RMSE}^{(s)}_{B,j,H} = \left[\frac{1}{N_H}\sum_i \left(e^{(s)}_{B,i,H,j}\right)^2\right]^{1/2},$$
$$r^{(s)}_{B,H} = \mathrm{median}_j\, \mathrm{RMSE}^{(s)}_{B,j,H},$$
$$R_H = \frac{1}{3}\sum_{s=1}^{3}\left(r^{(s)}_{A,H} - r^{(s)}_{E,H}\right).$$

In the two TD-MPC2 Cheetah checkpoints, R_H increases substantially over the tested horizon ladder. At $H = 25$, the differences are approximately 0.344 and 0.473 for checkpoints 52001 and 52002. DreamerV3 shows the same qualitative increase within its own native time scale. The numerical values should not be compared directly across systems because the queries and action repeats differ. Figure 5 shows the complete within-system horizon profiles.

C.2 Task, query, and rollout boundaries

The detailed Cheetah depth trajectory does not transfer unchanged to other task or protocol choices. In Finger Spin, the long-minus-short contrast is positive, but the trajectory starts negative and becomes more negative before crossing above zero at long horizon. Cup Catch also has a positive long-minus-short change, while the $H = 25$ aggregate remains close enough to zero that its interval crosses zero.

Changing the Cheetah physical query also changes the apparent short-horizon behavior. Under a harmonized $qpos + qvel$ target, the two TD-MPC2 checkpoints begin with a substantially negative $H = 1$ difference instead of reproducing the near-zero onset of the original diagnostic query. A controller-aligned TD-MPC2 evaluation at $H = 3$ produces much smaller and readout-dependent differences: nonlinear readouts are positive at both checkpoints, whereas affine intervals overlap zero.

Together, these comparisons show that the detailed horizon profile is a property of the task, query, readout, and rollout protocol rather than a universal signature of autonomous prediction. Table 7 summarizes the settings and the boundary each one establishes.

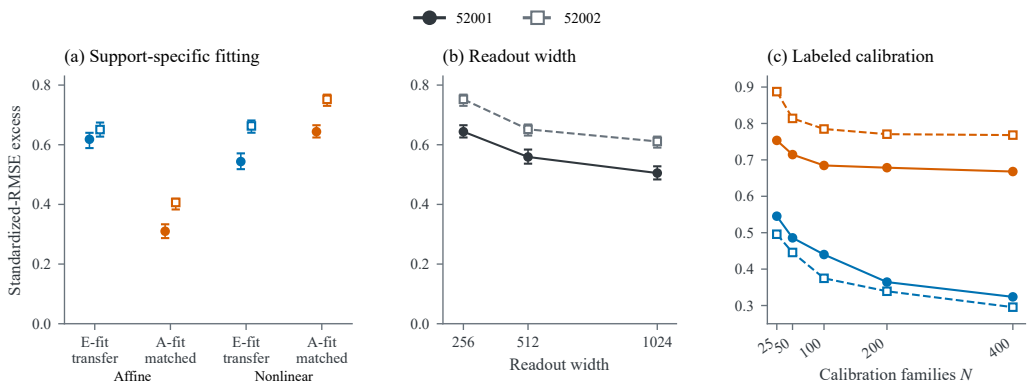

Figure 6: Readout controls for the measured E/A difference. Fitting the autonomous readout directly on autonomous features does not remove the difference; increasing readout width reduces it over the tested range; and additional labeled calibration improves both branches. Panel (c) shows separate branch errors (autonomous: orange; observation-aided: blue), not error contrasts. Because the largest readout and calibration settings have not saturated, these experiments do not support an irreducibility claim.

Table 7: Boundary settings for the depth and readout results. Each row retains its own task, target, and controller contract.

Setting	Changed factor	Main observation	Interpretation
Finger Spin	Task and native physical query	Negative short-horizon difference; positive long-minus-short contrast with a different intermediate trajectory.	A different depth profile, not a clean task replication.
Cup Catch	Task and native physical query	Positive long-minus-short contrast; the long-horizon interval includes zero.	Partial crossover only.
Harmonized Cheetah query	All qpos and qvel coordinates, retained without deletion	Short-horizon difference is negative at both checkpoints despite positive long-horizon differences.	The specified near-zero onset profile does not carry over to this query.
Controller-aligned Cheetah, $H = 3$	Native-policy action distribution and horizon	Affine readout intervals include zero; nonlinear readouts show small positive gaps.	Distinct from the fixed-action diagnostic at $H = 3$.

C.3 Readout support, width, and calibration

A readout trained only on observation-aided features could perform poorly on autonomous features because of support mismatch. We therefore fit g_A directly on autonomous features. The separately calibrated autonomous readout remains less accurate than the observation-aided readout in both TD-MPC2 checkpoints, so the observed contrast is not produced solely by transferring an E-trained readout to A.

Readout capacity nevertheless changes the measured difference. Increasing the nonlinear width from 256 to 512 to 1024 reduces the support-matched contrast on both checkpoints. Performance is still improving at the largest tested width, so these experiments do not identify an irreducible representational gap.

Additional labeled calibration data also improve both branches. Over the tested range, the observation-aided branch improves faster than the autonomous branch. We therefore interpret all reported physical errors as properties of the specified finite readout and calibration protocol rather than as information-theoretic limits of the latent state. Figure 6 collects the support, width, and calibration controls.

C.4 DECODE-AND-REENCODE CONTROL

We additionally test whether passing an autonomous latent state once through the model's own image decoder and observation encoder yields a representation from which physical state is easier to read.

Let A and E denote the original $H = 25$ representations and S_A, S_E their decoded-and-reencoded counterparts. Separate readouts of the same architecture are fitted on the recoded supports. At $H = 25$, $(L_E, L_A, L_{S_E}, L_{S_A}) = (0.230147, 0.463661, 0.236059, 0.466684)$. Define the autonomous recoding improvement as $\Delta_A = L_A - L_{S_A}$, so positive values indicate lower error after recoding. Here $\Delta_A = -0.003023$, with 95% CI $[-0.006971, -0.000671]$. For comparison, $\Delta_E = L_E - L_{S_E} = -0.005912$. The difference between the two recoding effects is 0.002889, with 95% CI $[-0.001696, 0.005814]$.

The tested decode-and-reencode transformation therefore does not improve autonomous physical readout in this setting.

C.5 ENDPOINT VERSUS PATH-LEVEL VISUAL ERROR

The main paper uses full-path visual loss rather than endpoint-only reconstruction error. On the DreamerV3 development bank, endpoint error has a root-level Spearman association with D_i of approximately 0.216, whereas the 25-step path loss has an association of approximately 0.261. The two lowest-loss quintiles overlap only partially: 252 of the 400 roots are common to both definitions. Full-path loss therefore incorporates temporal prediction error absent from a final-frame comparison and is used consistently in the main conditional analyses.

D FINITE-SCALE PHYSICAL SENSITIVITY

Section 4 asks why visually similar autonomous rollouts can have different physical errors. This appendix defines the finite-scale sensitivity calculation, its aggregation across perturbation scales, and its validation under actual environment perturbations.

D.1 PERTURBATION SETUP

Let x_i denote the environment state associated with root i. Starting from x_i, we replay the same action sequence after symmetric perturbations at scale h. Let $\Phi_O(x_i)$ denote the rendered rollout under the fixed image normalization and $\Phi_Q(x_i)$ the standardized physical query at the final horizon.

For each of the 18 root coordinates, the perturbation is $\pm h$ in calibration-standardized units, followed by the same fixed 25-action rollout. Central responses divide the difference between the two resulting outputs by $2h$. The observation response $J^O_{i,h}$ uses all 25 post-action $64 \times 64 \times 3$ RGB frames scaled by $1/255$; the query response $J^Q_{i,h}$ uses the 18 final $qpos + qvel$ coordinates divided by their calibration standard deviations. The executed forms are $G^O_{i,h} = (J^O_{i,h})^\top J^O_{i,h}/(25 \cdot 64 \cdot 64 \cdot 3)$ and $G^Q_{i,h} = (J^Q_{i,h})^\top J^Q_{i,h}/18$.

D.2 RELATIVE QUERY SENSITIVITY

For each root and perturbation scale,

$$\gamma_{i,h} = \max_{v \neq 0} \frac{v^\top G^Q_{i,h} v}{v^\top \left(G^O_{i,h} + \lambda I\right) v},$$

with fixed ridge $\lambda = 10^{-6}$. Equivalently, $\gamma_{i,h}$ is the leading generalized eigenvalue of

$$G^Q_{i,h} v = \gamma \left(G^O_{i,h} + \lambda I\right) v.$$

A large $\gamma_{i,h}$ indicates a finite perturbation direction along which the physical query changes strongly relative to the rendered trajectory. We do not interpret either the magnitude or the maximizing direction as scale invariant.

D.3 MULTI-SCALE SCORE

We use four fixed perturbation scales, $h \in \{0.0025, 0.005, 0.01, 0.02\}$. For each h, let F_h^{dev} be the empirical CDF of $\log \gamma_{i,h}$ on the development roots. For a new root,

$$u_i(h) = F_h^{\mathrm{dev}}(\log \gamma_{i,h}),$$

and

$$S_i = \mathrm{median}_h \, u_i(h).$$

The percentile mapping removes scale-specific magnitude differences before aggregation. The median prevents any one scale from determining the score.

D.4 SCALE BEHAVIOR ACROSS COHORTS

The rank correlations between neighboring scales are:

	.0025 → .005	.005 → .01	.01 → .02
Development	0.8650	0.8875	0.7265
Cohort 1	0.7896	0.7616	0.8366
Cohort 2	0.7610	0.7706	0.6625

The full four-scale ordering is substantially more stable than the weakest pairwise relation. The minimum leave-one-scale-out correlation with the full internal score is approximately 0.9843 in Development, 0.9632 in Cohort 1, and 0.9552 in Cohort 2. In Cohort 2, the score computed using the development CDFs agrees with the aggregate ordering within the cohort at $\rho = 0.9947$.

A family-cluster bootstrap for the weaker .01 → .02 pair gives a descriptive 95% interval of approximately

$$[0.542, 0.763].$$

The weaker relation is therefore retained as a finite-scale boundary rather than removed. Figure 7 provides the full observation-geometry, subspace, and leave-one-scale-out diagnostics.

D.5 NONLINEAR INTERVENTION VALIDATION

For each root and scale, the leading direction is the eigenvector for the largest eigenvalue returned by the symmetric-definite generalized eigensolver for $(G_{i,h}^Q, G_{i,h}^O + 10^{-6} I)$. We divide that vector by its Euclidean norm and orient it so its largest-magnitude component is positive. This normalization is applied in the 18 calibration-standardized root $qpos + qvel$ coordinates, before mapping the displacement to physical state units. In the following display, x_i denotes the 18 root $qpos + qvel$ coordinates in calibration-standardized units; all other full-state components are held fixed:

$$x_i^{\pm} = x_i \pm h v_{i,h}^{\star}.$$

After restoring the same unperturbed full root state, the physical displacement is $\pm h(\sigma^{\mathrm{state}} \odot v_{i,h}^{\star})$: its first nine entries are added to $qpos$ and its last nine to $qvel$. Each signed intervention replays the same 25 actions. The reference is the unperturbed rollout from that root under those actions, with RGB frames $O_{i,0}$ and final query $q_{i,0}$. For the perturbed rollout $(O_{i,\pm}, q_{i,\pm})$, we compute squared responses as follows.

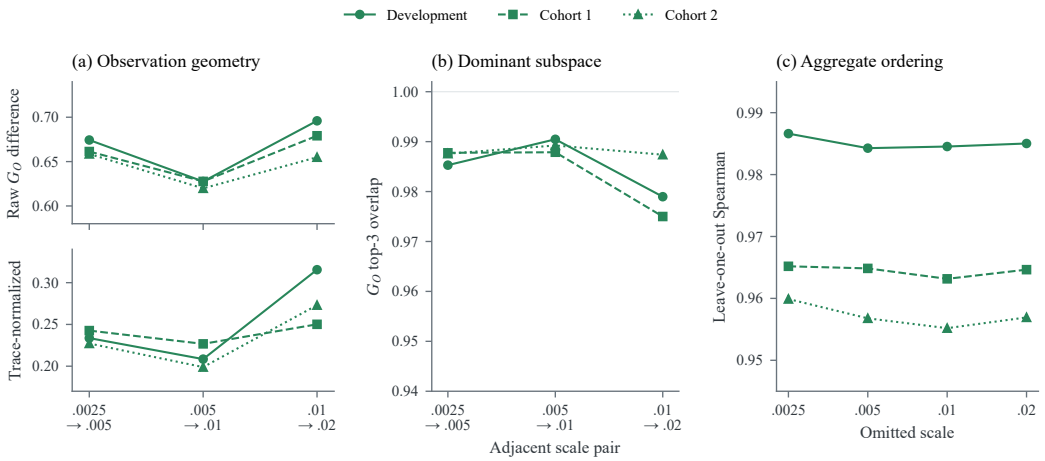

Figure 7: Detailed finite-scale diagnostics. Raw visual sensitivity differences are larger than trace-normalized observation-geometry differences, while dominant observation subspaces and aggregate root ordering remain comparatively stable. In all three cohorts, omitting any one scale leaves the four-scale ordering highly correlated with the full internal score. The weaker agreement between the two largest scales in Cohort 2 is retained as a finite-scale boundary.

$$Q^{\text{actual}}_{i,h,\pm} = \frac{1}{18}\left\|\frac{q_{i,\pm} - q_{i,0}}{\sigma^{\text{query}}}\right\|_2^2, \qquad V^{\text{actual}}_{i,h,\pm} = \frac{1}{25 \cdot 64 \cdot 64 \cdot 3}\left\|\frac{O_{i,\pm} - O_{i,0}}{255}\right\|_F^2.$$

Here O contains all 25 post-action $64 \times 64 \times 3$ RGB frames, and σ^{state} and σ^{query} are the coordinatewise calibration standard deviations for root state and final query. The executed analysis averages the two signed ratios, each with 10^{-12} added to its visual-response denominator:

$$R^{\text{actual}}_{i,h} = \frac{1}{2}\sum_{\pm}\frac{Q^{\text{actual}}_{i,h,\pm}}{V^{\text{actual}}_{i,h,\pm} + 10^{-12}}.$$

At each scale, we compute the Spearman association between $\log \gamma_{i,h}$ and $\log R^{\text{actual}}_{i,h}$ and average the four scale-specific correlations to obtain T.

The nonlinear intervention relationship is positive in both independent cohorts. In Cohort 2, $T = 0.4097$ with 95% CI $[0.3000, 0.5070]$.

The complete scale-specific intervention statistics are provided in the machine-readable supplementary table generated from the cohort analyses.

D.6 Regularization check

The generalized-eigenvalue calculation was repeated under the original ridge, a ten-fold decrease, a ten-fold increase, and a trace-relative alternative. The resulting root rankings were unchanged across these numerical checks. We therefore treat the ridge battery as a numerical check rather than as evidence for a broader robustness claim.

E Independent Prediction, Prospective Selection, and Additional Controls

This appendix gives the complete predictive analysis behind Section 4.3 and the controls used to delimit its interpretation.

E.1 SECOND INDEPENDENT COHORT AND ANALYSIS ORDER

The sensitivity score is computed without using physical readout errors from either independent cohort. Cohort 1 contains 96 roots from 67 episode families. Cohort 2 contains 160 roots from 117 new families.

Before physical readout outcomes were examined in Cohort 2, the finite-scale analysis found one weaker relationship between neighboring scales, $\rho_\gamma(.01, .02) = 0.6625$.

The four-scale score itself nevertheless remained highly coherent: every leave-one-scale-out correlation exceeded 0.955, and the score computed using the development CDFs agreed with the aggregate ordering within the cohort at $\rho = 0.9947$. The weaker pairwise relation was therefore retained as a scale boundary. Physical readout errors were examined only after the score and nonlinear intervention analysis were complete. Table 8 summarizes the order of analyses in Cohort 2.

Table 8: Order of analyses in Cohort 2.

Stage	Information used	Result	d_A, d_E, D accessed?
four-scale structure	finite perturbation geometry	one adjacent pair weaker	no
aggregate score	same geometry	leave-one-scale-out > 0.955	no
nonlinear intervention	environment responses	positive relation	no
physical prediction	fixed S, P^{path}, readouts	primary d_A test	yes

E.2 PARTIAL RANK CORRELATION

For each cohort, let $X_i = \mathrm{rank}(S_i)$, $P_i = \mathrm{rank}(P_i^{\mathrm{path}})$, and $Y_i = \mathrm{rank}(d_{A,i})$. We residualize the first and third ranks on the visual-loss rank,

$$X_i = \alpha_X + \beta_X P_i + \varepsilon_i^X, \qquad Y_i = \alpha_Y + \beta_Y P_i + \varepsilon_i^Y, \qquad \rho_A = \mathrm{corr}(\varepsilon^X, \varepsilon^Y).$$

Within every bootstrap replicate over episode families, the ranks and both residualizations are recomputed.

E.3 INDEPENDENT PREDICTION OF AUTONOMOUS ERROR

Cohort 1 gives $\rho_A = 0.2054$ with 95% CI $[0.0058, 0.3835]$, while Cohort 2 gives $\rho_A = 0.1631$ with 95% CI $[0.0131, 0.3122]$.

The two cohorts are reported separately; no pooled estimate is used to establish the result. Table 9 gives the cohort-specific prediction results alongside the intervention summary.

Table 9: Cohort-specific autonomous-error prediction.

Cohort	Roots	Families	ρ_A	95% CI	Intervention
Cohort 1	96	67	0.2054	[0.0058, 0.3835]	positive
Cohort 2	160	117	0.1631	[0.0131, 0.3122]	$T = 0.4097$ [0.3000,0.5070]

E.4 OBSERVATION-AIDED ERROR AND THE E/A CONTRAST

The same residual-rank procedure is applied to d_E and to $D = d_A - d_E$. The association with D is unsupported in both cohorts. The sensitivity score should therefore not be interpreted as an explanation of the autonomous-minus-observation-aided gap. Exact cohort-specific values and bootstrap intervals for these secondary analyses are reported in the supplementary machine-readable table.

E.5 SIMPLE CONTROLS FOR TRAJECTORY DIFFICULTY

A remaining alternative is that S marks generally difficult or highly dynamic trajectories. We therefore condition on three prespecified, directly available proxies.

The three controls are

$$B_{1,i} = \left\| \widetilde{q}_{i,0}^{\mathrm{vel}} \right\|_2,$$
$$B_{2,i} = \left(\frac{1}{H d_a} \sum_{h=1}^{H} \sum_{k=1}^{d_a} a_{i,h,k}^2 \right)^{1/2},$$
$$B_{4,i} = \frac{1}{H-1} \sum_{h=1}^{H-1} \mathrm{MSE}(o_{i,h+1}, o_{i,h}).$$

A physical-displacement proxy was not used because the evaluated $qpos$ variables did not have a single pre-existing wrap-aware difference convention suitable for this control. A support-distance proxy was also omitted because there was no canonical stored support measure to reuse without defining a new quantity after the primary result.

After additionally conditioning on B_1, B_2, B_4, Cohort 1 gives

$$\rho_{\mathrm{ext}} = 0.1873, \qquad 95\%\ \mathrm{CI} = [-0.0216, 0.3837],$$

retaining 91.2% of the original point estimate. Cohort 2 gives

$$\rho_{\mathrm{ext}} = 0.1547, \qquad 95\%\ \mathrm{CI} = [-0.0029, 0.3096],$$

retaining 94.9% of the original point estimate. The point estimates change little, while the wider intervals include zero. The analysis therefore argues against these simple proxies numerically absorbing the original association, but it is not a second independently significant conditional result. Table 10 reports both cohorts.

Table 10: Simple controls for trajectory difficulty.

Cohort	Original ρ_A	Extended ρ	95% CI	Retained
Cohort 1	0.2054	0.1873	[-0.0216, 0.3837]	91.2%
Cohort 2	0.1631	0.1547	[-0.0029, 0.3096]	94.9%

E.6 PROSPECTIVE TWO-STAGE SELECTION

Fresh Cohort 3 has 1,000 analysis roots from 500 episode families. Before accessing d_A, we fixed two-stage selection of the lowest-P^{path} 20% followed by the lowest-S half, versus direct lowest-P^{path} 10%. Both retain 100 roots; S uses the unchanged development-defined mappings of Eq. 2.

The selection contrast is

$$\Delta_{\mathrm{sel}} = \bar{d}_A^{\mathrm{two-stage}} - \bar{d}_A^{\mathrm{visual}},$$

Negative values favor sensitivity refinement. The prespecified minimum relative reduction was 5%; the 95% interval uses 10,000 whole-family bootstrap replicates. For each resampled bank, both selection rules are re-applied before computing their paired mean-error contrast; the development percentile maps remain fixed. Ties in P^{path} or S are broken by lexicographic root ID and then by sampled-family occurrence index. Table 11 reports the selected means.

Table 11: Prospective two-stage selection at matched 10% final coverage.

Policy	Selection rule	Roots	Mean d_A
visual-only	lowest P^{path} 10%	100	0.19786
two-stage	lowest P^{path} 20% $\rightarrow$ lowest-S half	100	0.16626

The result is $\Delta_{\mathrm{sel}} = -0.03160$ (95% CI $[-0.04881, -0.01616]$), a 15.97% relative reduction. Post-confirmatory descriptive checks give the same direction for all three readout fits (-0.0309, -0.0356, -0.0283); the visual-only and two-stage sets share 54 of 100 roots.

F Zero-Loss Certification: Statements and Proofs

The empirical experiments observe only positive visual losses. This appendix gives the full mathematical statements behind Section 5.

F.1 Minimal certificate

Let (Γ, d) be a compact metric space and let

$$V, R : \Gamma \to [0, \infty)$$

be continuous functions with

$$\min_{x \in \Gamma} V(x) = 0.$$

For $t \geq 0$, define

$$\Omega(t) = \max \{R(x) : V(x) \leq t\}.$$

Proposition F.1. The function Ω is finite and nondecreasing, and

$$R(x) \leq \Omega(V(x))$$

for every $x \in \Gamma$.

Moreover, the following are equivalent:

$$\Omega(t) \to 0 \quad \text{as } t \downarrow 0,$$

$$V^{-1}(0) \subseteq R^{-1}(0),$$

and

$$V(x_n) \to 0 \quad \Longrightarrow \quad R(x_n) \to 0$$

for every sequence $(x_n) \subset \Gamma$.

Finally, Ω is the smallest nondecreasing certificate: if ϕ is nondecreasing and

$$R(x) \leq \phi(V(x))$$

for every x, then

$$\Omega(t) \le \phi(t)$$

for every t.

Proof. Finiteness and monotonicity follow from compactness and nestedness of the sublevel sets. For any x,

$$x \in \{y : V(y) \le V(x)\},$$

so

$$R(x) \le \Omega(V(x)).$$

Suppose $V^{-1}(0) \subseteq R^{-1}(0)$ but $\Omega(t) \not\to 0$. Then there exist $\delta > 0$, $t_n \downarrow 0$, and x_n with

$$V(x_n) \le t_n, \qquad R(x_n) \ge \delta.$$

Compactness gives a convergent subsequence $x_{n_k} \to x^\star$. Continuity implies $V(x^\star) = 0$ while $R(x^\star) \ge \delta$, contradicting the zero-set inclusion.

Conversely, if $\Omega(t) \to 0$ and $V(x) = 0$, then for every $t > 0$,

$$R(x) \le \Omega(t),$$

hence $R(x) = 0$. The sequential formulation is equivalent by the same compactness argument.

Finally, if ϕ is nondecreasing and $R \le \phi \circ V$, then for every x with $V(x) \le t$,

$$R(x) \le \phi(V(x)) \le \phi(t).$$

Taking the maximum over the sublevel set yields $\Omega(t) \le \phi(t)$. □

F.2 Positive-loss screening and the zero-loss limit

Theorem F.2. Fix $0 < \varepsilon < 1$ and $0 < c < 1$. Let $\Gamma = [0, 1]$ and $V(x) = x$

with the uniform probability measure. There exist two nonnegative C^1 target-error functions R_{cert} and R_{fail} such that:

1. $R_{\mathrm{cert}}(x) = R_{\mathrm{fail}}(x)$ for all $x \ge \varepsilon$;
2. for every $t \in [\varepsilon, 1]$, the selected mass and selected mean target error are identical:

   $$\Pr[V \le t] = t,$$

   and

   $$\mathbb{E}[R_{\mathrm{cert}} \mid V \le t] = \mathbb{E}[R_{\mathrm{fail}} \mid V \le t] = \frac{t}{2} + \frac{c\varepsilon}{3t};$$

3. for every $t \in [\varepsilon, 1)$, lower-loss selection improves mean target error relative to the full population:

   $$\mathbb{E}[R] - \mathbb{E}[R \mid V \le t] = \frac{(1-t)(3t - 2c\varepsilon)}{6t} > 0;$$

4. R_{cert} admits a vanishing certificate whereas R_{fail} does not.

Thus the complete curve of mean error over selected low-loss sets above the positive loss floor ε does not identify whether a vanishing zero-loss certificate exists.

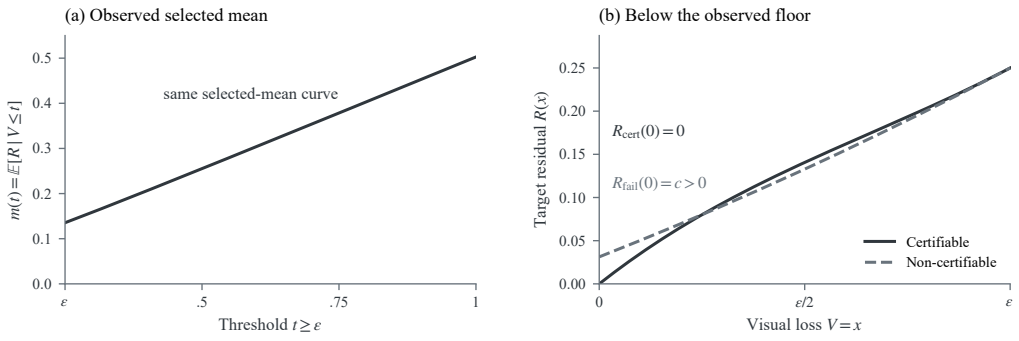

Figure 8: Explicit construction used in Theorem F.2. The two target-error constructions select the same mass and have the same mean target error at every observed threshold $t \geq \varepsilon$, but differ at $V = 0$. The plotted parameters are illustrative.

The restriction $0 < \varepsilon < 1$ in Theorem F.2 is only a normalization. Given any desired loss floor $\varepsilon_0 > 0$, apply Theorem F.2 with normalized floor $1/2$ on $[0, 1]$, and write $a = 2\varepsilon_0$. On $\Gamma_0 = [0, a]$ with the uniform probability measure, set $V_0(x) = x$ and $R_{j,0}(x) = aR_j(x/a)$ for $j \in \{\text{cert}, \text{fail}\}$. Then $x = \varepsilon_0$ corresponds to the normalized floor $1/2$; selected masses and conditional means above the floor are the rescaled versions of those in Theorem F.2, strict screening improvement holds for $t \in [\varepsilon_0, a)$, and the certifiable/non-certifiable zero-loss behavior is preserved. Moreover, $R'_{j,0}(x) = R'_j(x/a)$, so the C^1 and common-Lipschitz properties are unchanged. This gives the main-text statement for an arbitrary positive loss floor.

F.3 Explicit construction

For $0 \leq x \leq \varepsilon$, let

$$u = \frac{x}{\varepsilon},$$

and define

$$R_{\text{cert}}(x) = x + 4c\,u(1-u)^2,$$

$$R_{\text{fail}}(x) = x + c(1-u)^2.$$

For $x \geq \varepsilon$, set

$$R_{\text{cert}}(x) = R_{\text{fail}}(x) = x.$$

Both functions are nonnegative. At $x = \varepsilon$, $u = 1$, so the added terms and their first derivatives vanish. The functions therefore join $R(x) = x$ continuously with matching derivative and are C^1 on $[0, 1]$.

At zero, $R_{\text{cert}}(0) = 0$ whereas $R_{\text{fail}}(0) = c$.

Figure 8 visualizes the shared observed mean-error curve and the different behavior below the observed loss floor.

F.4 Why the observed means agree

For $t \geq \varepsilon$,

$$\int_0^t R_{\mathrm{cert}}(x)\,dx = \frac{t^2}{2} + 4c\varepsilon \int_0^1 u(1-u)^2\,du.$$

Since

$$\int_0^1 u(1-u)^2\,du = \frac{1}{12},$$

we obtain

$$\int_0^t R_{\mathrm{cert}}(x)\,dx = \frac{t^2}{2} + \frac{c\varepsilon}{3}.$$

Similarly,

$$\int_0^t R_{\mathrm{fail}}(x)\,dx = \frac{t^2}{2} + c\varepsilon \int_0^1 (1-u)^2\,du,$$

and

$$\int_0^1 (1-u)^2\,du = \frac{1}{3}.$$

Thus

$$\int_0^t R_{\mathrm{fail}}(x)\,dx = \frac{t^2}{2} + \frac{c\varepsilon}{3},$$

and because the selected mass is t,

$$m(t) = \mathbb{E}[R \mid V \le t] = \frac{t}{2} + \frac{c\varepsilon}{3t}.$$

F.5 Selection remains useful

The mean over the full population is

$$m(1) = \frac{1}{2} + \frac{c\varepsilon}{3}.$$

Therefore

$$m(1) - m(t) = \frac{(1-t)(3t - 2c\varepsilon)}{6t}.$$

For $t \ge \varepsilon$ and $0 < c < 1$,

$$3t - 2c\varepsilon \ge (3 - 2c)\varepsilon > 0.$$

Hence $m(1) > m(t)$ for every $t < 1$ in the observed range. The construction therefore does not rely on an uninformative visual metric: low-loss selection is genuinely useful while zero-loss certification remains unidentified.

F.6 Different zero-loss behavior

For the certifiable construction, $R_{\mathrm{cert}}(0) = 0$, and continuity implies

$$\sup_{x \le t} R_{\mathrm{cert}}(x) \to 0$$

as $t \downarrow 0$.

For the non-certifiable construction,

$$R_{\mathrm{fail}}(0) = c,$$

so for every $t > 0$,

$$\Omega_{\mathrm{fail}}(t) = \sup_{x \le t} R_{\mathrm{fail}}(x) \ge c.$$

No vanishing certificate can therefore exist. This proves Theorem F.2. □

F.7 A common Lipschitz class

The separation persists under a shared finite Lipschitz bound. Let

$$c = a\varepsilon,$$

with

$$0 < a < \min\left\{\frac{1}{2}, \frac{L-1}{4}\right\}$$

for any prescribed $L > 1$.

On $x \le \varepsilon$,

$$R'_{\mathrm{fail}}(x) = 1 - 2a(1-u),$$

so $|R'_{\mathrm{fail}}(x)| \le 1$. For the certifiable construction,

$$R'_{\mathrm{cert}}(x) = 1 + 4a(1-u)(1-3u).$$

Since

$$-\frac{1}{3} \le (1-u)(1-3u) \le 1$$

on $u \in [0,1]$,

$$|R'_{\mathrm{cert}}(x)| \le 1 + 4a < L.$$

For $x \ge \varepsilon$, both derivatives equal 1. The two constructions can therefore be chosen within a common Lipschitz class while retaining the same screening curve above the observed loss floor and different behavior at zero loss.

G Implementation and Reproducibility

G.1 Experimental systems

The evaluation reuses fixed pretrained world-model checkpoints. TD-MPC2 contributes two Cheetah checkpoints, 52001 and 52002, for horizon and readout controls. DreamerV3 contributes the Cheetah checkpoint used for the main visual-selection and sensitivity analyses. PlaNet provides the second RSSM implementation for the visual-selection replication. No world-model parameters are updated during the evaluations reported in the paper.

G.2 Readout optimization

The main DreamerV3 and PlaNet physical readouts are two-layer GELU MLPs with hidden width 1024 and three independent fitting repetitions.

The TD-MPC2 width and calibration controls use AdamW with learning rate 3×10^{-4}, weight decay 10^{-4}, batch size 256, and fitting seeds 91001, 91002, and 91003. The width experiment evaluates hidden widths 256, 512, and 1024. The calibration experiment uses family-level fitting and validation splits with validation-selected early stopping. Because the two sweeps use different fitting and normalization protocols, their nominal width-1024 cells are not interpreted as the same experimental condition.

G.3 TD-MPC2 diagnostic query and actions

The frozen TD-MPC2 diagnostic query contains ten physical coordinates: `qpos.rootz`, `qpos.rooty`, the back- and front-leg thigh/shin/foot coordinates, `torso.height`, and `torso.upright`.

Hinge-angle coordinates use wrapped differences

$$((\theta + \pi) \bmod 2\pi) - \pi.$$

The main TD-MPC2 horizon ladder uses a fixed counterfactual action stream consisting of 25 six-dimensional normalized actions. This stream is diagnostic rather than the executed future policy trajectory. The controller-aligned $H = 3$ experiment in Appendix C uses a separate native protocol and is not treated as the $H = 3$ point of the diagnostic ladder.

G.4 DreamerV3 and PlaNet evaluation sets

DreamerV3 development and independent visual-selection banks each contain 2,000 roots from 100 source episodes. PlaNet uses pairwise-disjoint calibration, development, and independent sets, each containing 2,000 roots from 100 source episodes. The sensitivity development set, Cohort 1, Cohort 2, and Cohort 3 use separate root/family collections as summarized in Table 3. The supplementary package provides anonymized analysis-level source tables for the reported visual-selection, finite-scale, prospective-selection, and Appendix C analyses.

G.5 Perturbation implementation

The finite-scale analysis uses $h \in \{0.0025, 0.005, 0.01, 0.02\}$ and ridge $\lambda = 10^{-6}$ in the generalized eigenvalue problem. All scales reuse the same query definition, image normalization, rollout horizon, and action sequence for a given root. The development empirical CDFs F_h^{dev} are stored and reused unchanged for both independent cohorts; independent cohorts do not define new percentile maps.

Before interpreting the second independent-cohort perturbation calculations, an exact replay check reproduced the final $qpos + qvel$ endpoint on all 16 sampled roots used for that audit.

G.6 Random seeds and resampling

All inferential confidence intervals use episode-family resampling. The supplementary analysis code contains the fixed seeds and resampling procedures needed to reproduce the reported intervals. The final simple-proxy control uses dedicated bootstrap seeds 26092551 and 26092552 for Cohort 1 and Cohort 2, respectively.

G.7 Figure and table source data

All empirical manuscript figures and tables are generated from machine-readable analysis-level sources or fixed table sources. The supplementary package contains the corresponding anonymized inputs and standalone analysis, plotting, and table-generation code. Figure 1 is an authored schematic and is supplied as a vector asset.

G.8 Supplementary artifacts

The supplementary package is designed for paper-level reproduction. It includes anonymized analysis-level data and standalone scripts for reproducing the reported statistics, figures, and tables. Model training, rollout collection, checkpoints, and general experiment infrastructure are outside the scope of the package and are not required for the included reproduction.