# THE CORRECTION BUDGET: WHEN TEST-TIME REFINEMENT CAN AND CANNOT FIX WORLD-MODEL ROLLOUTS

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

## ABSTRACT

Test-time refinement, which predicts a latent and then projects it onto a learned constraint, is an appealing recipe against the compounding drift of action-conditioned world models. We build the recipe in its strongest practical form (a transition energy model trained against the world model's own drifted rollouts, fixing verifier area-under-the-receiver-operating-characteristic-curve (AUROC) from 0.06 to 0.88–0.94) and then take it apart. Across three environments, two observation modalities, and fifteen (world-model, corrector) systems, we find: (i) drift does not compound exponentially; the off-manifold amplification premise fails; (ii) the verifier is a pure *legality* detector: its gradients, regression targets, and selection weights carry no information about the direction back to the truth; (iii) a maximal-input probe (the world model fine-tuned on its own rollout contexts) shows that open-loop error has a predictable fraction governed by two separable channels, reducible model bias (data starvation) and off-manifold geometry (representation); (iv) a correction-budget law: no test-time corrector measurable in the rollout's inputs can beat $\delta_{\text{best}}/\delta_{\text{raw}} \geq \sqrt{\text{FVU}^*}$, where FVU is the fraction of variance unexplained. The bound is saturated exactly where per-step alignment is high (pixel latents, where legality $\approx$ truth and refinement recovers up to 46% of drift) and correctly forecasts the near-zero gains on near-manifold proprioception.

## 1 INTRODUCTION

Predict, then project. A generator proposes; a learned verifier checks or corrects. The recipe powers process reward models and best-of-$N$ selection for language models, energy-based guidance for compositional generation, and safety projections for diffusion policies; it reads as the natural next step for action-conditioned world models: open-loop imagination drifts off the data manifold, drift is the known disease, and a learned check of physical legality is the advertised cure. This paper builds that cure in its strongest practical form, and then takes it apart.

Taking it apart required first making it work. A transition verifier trained the standard way, with negatives corrupted around ground-truth states, is not merely weak at test time, it is *inverted*: conditioned on the world model's own drifted rollouts it rates the model's hallucination as more legal than the truth (area under the receiver operating characteristic curve, AUROC 0.062). We trace the failure to a train/test conditioning mismatch and repair it with *paired-drift negatives*: condition on the world model's own open-loop predictions and make the hardest negative the model's self-consistent next step. Held-out drift-conditioned AUROC rises to 0.81–0.94 on every domain (Section 3). Everything dissected below is this healthy verifier, not a strawman.

What the dissection finds is not what the recipe promises. First, the motivating premise fails: open-loop drift is *linear* in horizon, not exponential; the fitted per-step amplification is $\hat{L} \approx 1.00$ on every reasonably-trained system, out to $T = 200$ (Section 3.1). Second, legality is not truth: the three and only three ways to extract a correction from a scalar energy (descend its gradient, regress toward it, select with it) all improve legality while leaving fidelity unchanged or worse, even at AUROC 0.94 (Section 4). Third, the failure is not the verifier's fault: a maximal-input probe (the world model fine-tuned on its own rollout contexts, an upper bound on what any test-time corrector could know)



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.

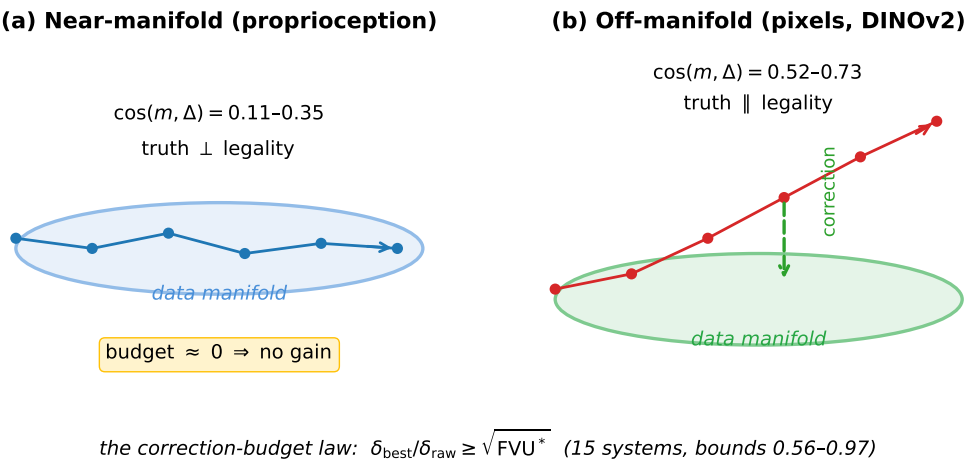



Figure 1: Two regimes of test-time correction. **(a)** Near-manifold rollouts (proprioception): truth direction $\perp$ legality direction ($\cos(m, \Delta) = 0.11$–$0.35$), budget $\approx 0$, refinement cannot pay. **(b)** Off-manifold rollouts (pixels, DINOv2): truth $\parallel$ legality ($\cos(m, \Delta) = 0.52$–$0.73$), budget is large, refinement pays up to the law.

shows that most open-loop error is simply not predictable from the rollout's inputs on proprioceptive state, and largely *is* predictable on pixel latents (Section 5).

These observations assemble into a quantitative account. The predictable fraction is governed by two separable channels: reducible model bias, which scales with data starvation, and off-manifold geometry, measured by $\cos(m, \Delta)$, which captures how far the direction back to the truth coincides with the direction back to the data manifold. That dial is set by the observation modality (0.11–0.35 on proprioception vs. 0.52–0.73 on pixels, across three environments and two encoders), not by the dynamics, the encoder, or the data budget. And the predictable fraction is a budget: for any corrector measurable in the rollout's inputs, $\delta_{\text{best}}/\delta_{\text{raw}} \geq \sqrt{\text{FVU}^*}$, where FVU denotes the fraction of variance unexplained, a correction-budget law we formalize (Section 6) and verify on fifteen systems with paired confidence intervals.

The law has predictive force in both directions. It correctly forecasts the near-zero (often negative) gains of every refinement method on near-manifold proprioception, and it names the regime where refinement pays: on pixel latents test-time correction recovers 29–46% of open-loop drift, and even a zero-learning nearest-neighbor return rule recovers 22–31%. Fig. 1 previews the two regimes.

**Contributions.**

- **A falsification and a fix.** The exponential-compounding premise behind predict-then-project fails on all tested domains (drift is linear, $\hat{L} \approx 1.00$). We isolate why naive transition verifiers collapse (AUROC 0.06) and fix it with paired-drift negatives (0.88–0.94).
- **A three-modality dissociation.** Gradient descent, regression, and verifier-weighted selection all improve legality while leaving truth unchanged or worse; the verifier carries $\sim$zero tangential information, even at AUROC 0.94.
- **Two-channel predictability.** A maximal-input $\Delta$-probe measures the predictable fraction of open-loop error and decomposes it: a bias channel (data starvation) and a geometry channel ($\cos(m, \Delta)$), set by modality.
- **The correction-budget law.** $\delta_{\text{best}}/\delta_{\text{raw}} \geq \sqrt{\text{FVU}^*}$ for any corrector measurable in the rollout's inputs; verified on 15 systems (bounds 0.56–0.97), saturated where per-step cos is high.

## 2 Setup

One frozen stack is shared by every experiment (full reproduction details: Appendix C).

**Frozen world model $W_\theta$:** causal Transformer, context $H = 16$, deterministic; open-loop imagination $\hat{z}_{t+1} = W(\hat{z}_{t-H+1:t}, a)$. Latents standardized per-dim (zero mean, unit std); all step sizes in those units.

**Verifier $E_\psi(z_t, a_t, z')$:** transition energy model; refinement (PMP) = $K$ steps of projected gradient descent on $E_\psi$ with trust region $\rho = 3\sigma_{\text{resid}}$, gradient normalization, early stop.

**Protocol E1** (policy-independent imagination fidelity): $H$-frame ground-truth (GT) context + GT action sequence, $T = 50$ open-loop steps; $\delta(t) = \mathbb{E}\|\tilde{z}_t - z_t\|$ (mean of norms). No policy, no reward; fidelity only.

**Baselines/oracles:** B2 = input-noise-trained WM; BoN = step-level best-of-$N$ by verifier energy; O1 = teacher forcing (one-step noise floor); O2 = oracle direction toward GT at same trust radius (upper bound for any projector).

**Domains:** walker/cheetah on DeepMind Control (dm_control; proprio $d_z = 17$–$24$); pixels are V-D4RL medium encoded once by frozen DINOv2 Vision Transformer Large (ViT-L) classification token (CLS; $d_z = 1024$); humanoid pixels as third env. 15 systems total: 9 canonical stacks + ViT-S pair + data-axis pair + humanoid pair.

**Statistics:** proprio $\delta@50$ scans use $n = 1024$ (cross-system gaps wobble $\pm$3–5% at $n = 256$); pixels $n = 256$ (norms concentrate $\sim$0.5%). All gaps carry paired 95% confidence intervals (CIs; 10k bootstrap).

# 3 The premise fails, and the verifier can be fixed

## 3.1 Drift is linear, not exponential

The motivating story for predict-then-project is exponential error compounding: once a rollout leaves the data manifold, the world model's effective Lipschitz constant exceeds one and per-step errors amplify geometrically. We test the premise directly. For every domain we fit the open-loop error curve $\delta(t)$ to a linear shape $a \cdot t$ and a geometric shape $\epsilon(L^t - 1)/(L - 1)$, and compare fits.

**The premise fails everywhere.** Across all five domains $\hat{L} \approx 1.001$ and the linear fit wins (geometric $R^2 <$ linear $R^2$ in every case, both modalities). Drift is steady accumulation, not amplification: there is no runaway process for a projection to arrest. At $T = 200$, where even $L = 1.01$ would amplify drift $\times$7.3, all four reasonably-trained systems remain linear at 1000/1000 bootstrap replicates. The one exception is informative: the deliberately data-starved walker100 turns geometric at $T = 200$ ($\hat{L} = 1.081$, 95% CI [1.033, 1.088]) while remaining linear at $T \leq 50$; the equally-starved pixel WMs stay linear; super-linear compounding is a symptom of severe proprio data starvation, not of modality or horizon.

## 3.2 Why naive verifiers collapse, and the fix

The standard recipe trains $E_\psi(z_t, a_t, z')$ conditioning every negative on the ground-truth current state $z_t$. Deployed inside imagination, the conditioning state is the model's own drifted prediction $\tilde{z}_t$, an input the verifier has never seen. The failure is not graceful: evaluated on drift-conditioned pairs, the GT-conditioned verifier scores AUROC **0.062**, systematically *inverted*: it rates the world model's hallucinated next step as more legal than the ground truth.

The repair follows from the diagnosis: train the verifier on the states it will actually be conditioned on. We roll the frozen WM forward $k$ steps (mixed depths) and condition every training pair on the drifted prediction $\tilde{z}_t$; the hardest negative is the WM's own self-consistent next step $W(\tilde{z}_t, a_t)$; energy zero is pinned only on genuinely on-manifold ($k = 0$) transitions. Held-out drift-conditioned AUROC rises to 0.81–0.94 (Table 1). Two caveats: data volume matters (120 episodes $\rightarrow$ AUROC 0.265, pure memorization); and a fixed ranking is not a fixed correction direction, the subject of Section 4.

Table 1: Held-out drift-conditioned AUROC (positive vs. hardest negative).

| Domain | GT-cond. | Paired-drift |
|---|---|---|
| Walker proprio | 0.062 | 0.812 |
| Cheetah proprio | 0.062 | 0.847 |
| Walker pixels | 0.062 | 0.939 |
| Cheetah pixels | 0.062 | 0.905 |
| Humanoid pixels | 0.062 | 0.921 |

# 4 LEGALITY IS NOT TRUTH: A THREE-MODALITY DISSOCIATION

With ranking healthy (AUROC 0.81–0.94), we ask whether the verifier can steer a rollout back toward the truth. There are exactly three ways to extract a correction from a scalar energy: descend its gradient, regress a direction toward its minima, or use it to select among samples. We test all three.

**Gradient.** Along open-loop rollouts, $\cos(-\nabla E, z_{\text{gt}} - \hat{z})$, alignment between the verifier's descent direction and the true correction, is $+0.03$ (walker proprio) and $+0.008$ (cheetah); for scale, a *random* direction at $d = 17$ has mean $|\cos| = 0.24$. On pixels it is modestly positive: $+0.168$ (walker) and $+0.105$ (cheetah). End-to-end PMP agrees: on proprio, any $K \geq 1$ roughly doubles horizon-mean error; on pixels it helps slightly ($-3.5\%$, $K = 1$ optimal).

**Regression.** Train a directional head $g_\psi(\tilde{z}_t, a_t, z') \leftarrow z' - z_{\text{gt}}$ on paired-drift candidates. Alignment improves $0.03 \rightarrow 0.14$, a real $5\times$ gain, but far below the $\sim$0.4 break-even. The failure mode is instructive: exactly at the WM's own prediction, the learned field's magnitude collapses to $\sim$0.09 (target $\sim$1.0); the closed loop lands at $\delta@50 = 3.17$ vs. raw 2.96, worst at $K = 1$. The field's fixed point is the network's own next-state estimate $\approx$ the WM's answer.

**Selection.** Verifier-weighted particle imagination (sequential importance resampling, SIR, with potentials $\exp(-\lambda \cdot \text{zscore}(E))$; uniform = PETS-style control; oracle GT weights = ceiling). Legality improves monotonically in $\lambda$ everywhere (walker $k$-nearest neighbor (kNN) $1.94 \rightarrow 1.62$ against GT reference 1.47) while truth *degrades*. On pixels, the proprio 'pure coverage win' (uniform particles beating a single rollout by 44–51%) inverts catastrophically: $\sigma$-scale perturbations in 1024-d knock trajectories off-manifold (mean-$\delta@50$ 264 vs. raw 49.5).

**Verdict.** Three extraction modalities (gradient, regression, selection) return one verdict: the paired-drift verifier is a pure *legality* detector with $\sim$zero truth-direction information, in both regimes, even at AUROC 0.939. Whether legality helps truth is therefore not a property of the verifier; it is a property of the domain's geometry, which we measure next.

# 5 THE $\Delta$-PROBE: PREDICTABLE FRACTION AND TWO CHANNELS

## 5.1 A MAXIMAL-INPUT PROBE

The verifier failed for lack of *information*, not lack of accuracy. So give a corrector the most information any test-time corrector can have: the rollout's own recent past. The $\Delta$-probe is a copy of the frozen WM, fine-tuned on the WM's own open-loop rollout contexts (mixed depths $j = 0 \ldots 39$; the base WM is frozen, so the input distribution is static) to regress the ground-truth next state. Its per-depth alignment $\cos(\hat{\Delta}, \Delta)$ and fraction of variance unexplained (FVU) measure how much of open-loop error is predictable at all from the imagination's own inputs (Fig. 2).

## 5.2 PREDICTABILITY AND THE GEOMETRY DIAL

**Predictability: capped on proprio, dominant on pixels.** On proprio, even the WM's full context predicts only 14–20% of open-loop error variance (FVU floors 0.80/0.86; cos caps 0.41/0.24). On pixels the picture inverts: 69% (walker) / 52% (cheetah) of error variance is predictable (floors 0.31/0.49, caps 0.83/0.73); these 100-trajectory WMs are data-starved (val one-step MSE 14–17$\times$ train).

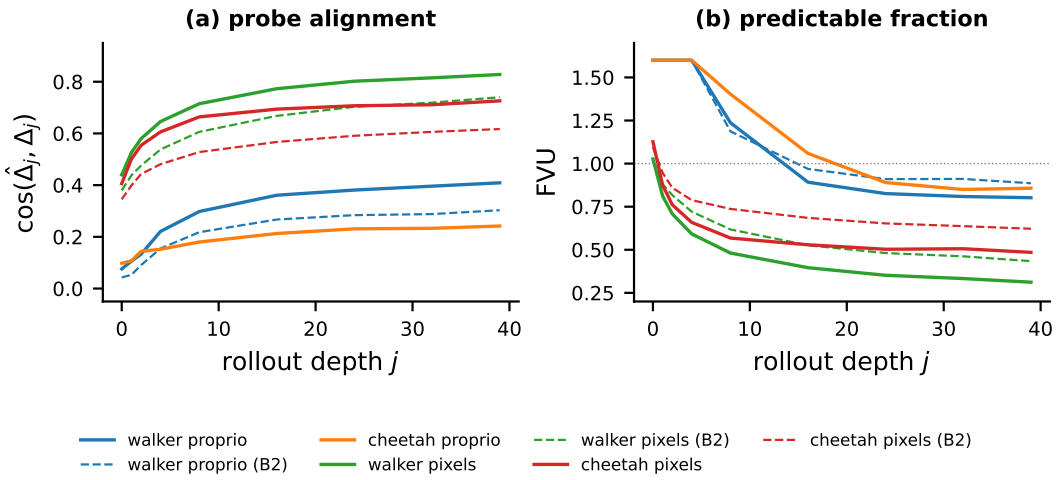



Figure 2: Probe predictability. **(a)** $\cos(\hat{\Delta}_j, \Delta_j)$ and **(b)** FVU vs. rollout depth. Solid = plain WM, dashed = B2. On proprio, even the WM's full context predicts only 14–20% of error variance (FVU floors 0.80–0.86); on pixels, 52–69% is predictable (floors 0.31–0.49).

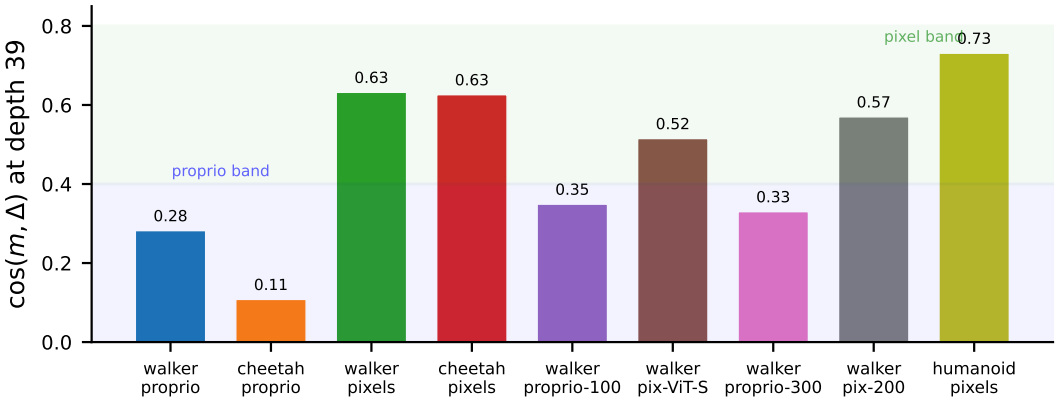



Figure 3: The geometry dial $\cos(m, \Delta)$ at depth 39 across all domains. Set by modality: 0.11–0.35 (proprio) vs. 0.52–0.73 (pixels), invariant to encoder (ViT-S), data budget, and environment.

**The geometry dial** $\cos(m, \Delta)$**.** What *kind* of error is the predictable part? Decompose the probe's output against the zero-learning legality direction $m = (\text{mean of 8 nearest training frames}) - \hat{z}$, the 'return to manifold' vector (Fig. 3).

Two facts (Tab. 2). First, on proprio the probe's signal is genuinely *tangential*: after projecting out the manifold-return direction, alignment persists ($\cos(\hat{\Delta}_\perp, \Delta_\perp) = 0.23$–$0.34$), and a zero-learning kNN-return rule fails in closed loop (walker 3.27, cheetah 1.64; both worse than raw). The hypothesis 'whatever is predictable is just legality' is refuted. Second, the dial $\cos(m, \Delta)$, which measures how much of the truth direction is simply 'return to the manifold', is set by the *modality*: 0.11/0.28 on proprio vs. 0.63/0.63 on pixels, nearly identical across environments within a modality. Replications keep the dial in its band: ViT-S 0.52, walker-pix-200 0.57, humanoid 0.73, proprio at 300/100 episodes 0.33/0.35, modality-locked under encoder swap, data budget, and a third environment.

### 5.3 Two channels: bias × geometry

Pixels differ from proprio in two ways at once: their WMs are data-starved (bias) and their rollouts leave the manifold (geometry). The ablation walker_walk_proprio100 (same environment, cache rebuilt with 100 episodes) separates the channels. Starvation alone raises the **bias** channel: the WM's val/train gap jumps $3.5\times \rightarrow 311\times$, the probe's cos cap rises $0.41 \rightarrow 0.47$, and closed-

Table 2: Decomposition at depth 39. $\cos(m, \Delta)$ is the geometry dial; $\cos(\hat{\Delta}_\perp, \Delta_\perp)$ after projecting out $m$.

| **Domain** | $\cos(m, \Delta)$ | $\cos(\hat{\Delta}_\perp, \Delta_\perp)$ | kNN closed-loop |
|---|---|---|---|
| Walker proprio | 0.28 | 0.34 | 3.27 (raw 2.96) |
| Cheetah proprio | 0.11 | 0.23 | 1.64 (raw 1.04) |
| Walker pixels | 0.63 | 0.43 | 1.47 (raw 2.68, $-45\%$) |
| Cheetah pixels | 0.63 | 0.38 | 1.40 (raw 1.96, $-29\%$) |
| Humanoid pixels | 0.73 | 0.40 | 1.25 (raw 2.03, $-38\%$) |

loop cash-in (Section 6) grows $-6.5\% \to -15.9\%$. But the **geometry** channel does not move: $\cos(m, \Delta)$ stays 0.35 (vs. 0.63 on pixels) and kNN-return still hurts. Data starvation makes more error predictable; only the observation modality makes the legality direction point at the truth.

## 6 THE CORRECTION-BUDGET LAW

**Statement (informal).** A closed-loop corrector is itself a function of the imagination's inputs; therefore its error is floored by the open-loop unpredictable fraction: $\delta_{\text{best}}/\delta_{\text{raw}} \geq \sqrt{\text{FVU}^*}$. The probe supplies a computable estimate from above (probe-FVU $\geq$ Bayes-FVU); oracles (O1/O2), which condition on the truth, are outside the class and must and do break it.

### 6.1 SETUP AND CORRECTOR CLASS

Write $\xi \triangleq (z_{-H+1:0}, a_{0:T-1})$ for the imagination inputs and $\mathcal{G} = \sigma(\xi)$. The frozen WM $W$ is deterministic, so the open-loop rollout $\hat{z}_t = W^{(t)}(\xi)$ is $\mathcal{G}$-measurable. Let $z_t$ be the ground-truth latent and $\Delta_t \triangleq z_t - \hat{z}_t$.

**Corrector class.** $\mathcal{C}$ is the class of all test-time imagination procedures whose horizon-$T$ output is $\tilde{z}_T = f(\xi, U)$, with $f$ measurable and $U$ algorithmic randomness independent of $(\xi, z_T)$. This covers PMP, particle populations, kNN-return, the probe-$\alpha$ corrector, and any iterated variant. Not in $\mathcal{C}$: O1/O2 and GT-selected particle minima, which condition on $z$ itself.

### 6.2 THE FLOOR

**Proposition 1** (Correction-budget floor)**.** *For every $\tilde{z}_T \in \mathcal{C}$, $\mathbb{E}\|\tilde{z}_T - z_T\|^2 \geq \mathbb{E}\|z_T - \mathbb{E}[z_T|\xi]\|^2$, with equality iff $\tilde{z}_T = \mathbb{E}[z_T|\xi]$ a.s. Consequently, in RMS units,*

$$\delta^{corr}_{RMS}(T)/\delta^{raw}_{RMS}(T) \geq \sqrt{FVU^*_T}, \quad FVU^*_T \triangleq \frac{\mathbb{E}\|z_T - \mathbb{E}[z_T|\xi]\|^2}{\mathbb{E}\|z_T - \hat{z}_T\|^2} \in [0, 1].$$

*Proof sketch.* Condition on $U = u$: $g_u \triangleq f(\cdot, u)$ is $\mathcal{G}$-measurable, and by the tower property the cross term in $\mathbb{E}\|g_u(\xi) - z_T\|^2 = \mathbb{E}\|g_u(\xi) - \mathbb{E}[z_T|\xi]\|^2 + \mathbb{E}\|\mathbb{E}[z_T|\xi] - z_T\|^2$ vanishes. Average over $u$. $\text{FVU}^*_T \leq 1$ because $\hat{z}_T$ is itself $\mathcal{G}$-measurable. □

### 6.3 A COMPUTABLE ESTIMATE FROM ABOVE

**Lemma 1** (Probe-FVU estimates the floor from above)**.** *For any probe $P$ with input $\chi_j = \varphi_j(\xi)$: $FVU^{probe}_j \geq FVU^*_{j+1}(\sigma(\chi_j)) \geq FVU^*_{j+1}(\mathcal{G})$.*

*Proof sketch.* First inequality: the conditional mean given $\chi_j$ minimizes MSE among $\sigma(\chi_j)$-measurable predictors, and the probe is one such predictor. Second: $\sigma(\chi_j) \subseteq \mathcal{G}$, and conditional variance is non-increasing under refinement of the conditioning $\sigma$-algebra. □

**Assumption A1** (depth monotonicity). $\text{FVU}^*_t(\mathcal{G})$ is non-increasing in $t$ beyond a burn-in depth $j_0$. Empirical support: the probe's FVU curve is flat or declining past depth $\approx 16$ in all 15 systems.

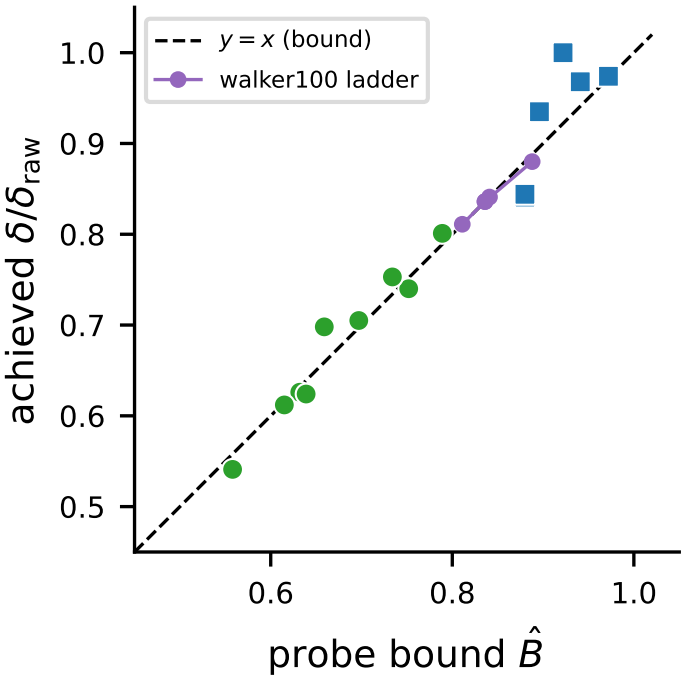



Figure 4: The correction-budget law: achieved $\delta$ ratio vs. probe-estimated bound $\hat{B}$ across 15 systems. ■ = proprio, ● = pixels. The $y = x$ line is the bound; saturation where per-step cos is high (pixels), unrealized allowance where it is low (proprio).

**Corollary 1** (The operational bound)**.** *Let* $\hat{B} \triangleq (\min_{j \le R-1} FVU_j^{probe})^{1/2}$ *(the plotted bound;* $R - 1 = 39$*). Under A1, for* $T \ge j^* + 1$*:* $\hat{B} \ge \sqrt{FVU^*_{j^*+1}(\mathcal{G})} \ge \sqrt{FVU^*_T(\mathcal{G})}$.

**Honest boundaries (R0–R5).** R0: corrector-induced distribution shift does not enter (the proof never decomposes step-wise). R1: the paper's empirical content is the *learnable* floor ($\hat{B}$ with the WM's own architecture estimates what is statistically extractable). R2: the theorem is in root-mean-square (RMS) units; measured $\delta$ is mean of norms; the mean-norm floor sits at or below the RMS floor, explaining sub-RMS achieved ratios. R3: iterated correctors remain in $\mathcal{C}$ but are not floored by the probe's own FVU. R4: each WM defines its own floor; B2's rollouts are less probe-predictable in 7/7 paired systems. R5: the law does not say refinement is useless (pixels: $-29\%$ to $-46\%$ realized, bound saturated). Detailed discussion in Appendix A.

**Empirical verification (Fig. 4).** The law holds on 14/15 systems in RMS units; the single below-line point (walker-pix-200, $-2.7$pp) is probe-limited (pre-registered ladder tightens $\hat{B}$ from 0.639 to 0.630). Three observations:

- **Saturation follows per-step cos.** Achieved $\approx \hat{B}$ exactly where per-step $\cos(\hat{\Delta}, \Delta)$ is high (all pixel systems, cos 0.62–0.83); in the low-cos regime the $1-\cos^2$ orthogonal component of every correction injects fresh error that compounds.
- **Information is spent once** (Fig. 5): on the corrected rollout the probe's alignment collapses within 2–4 steps. A per-step-skim simulator that re-draws the measured alignment every step over-predicts cash-in $3\times$, the double-counting negative control.
- **B2 as budget consumption.** Input-noise training consumes part of the same predictable stock; B2 rollouts are less probe-predictable in 7/7 systems; consumption $\not\Rightarrow$ realized gain (cheetah B2 consumes, gains $\sim$0).

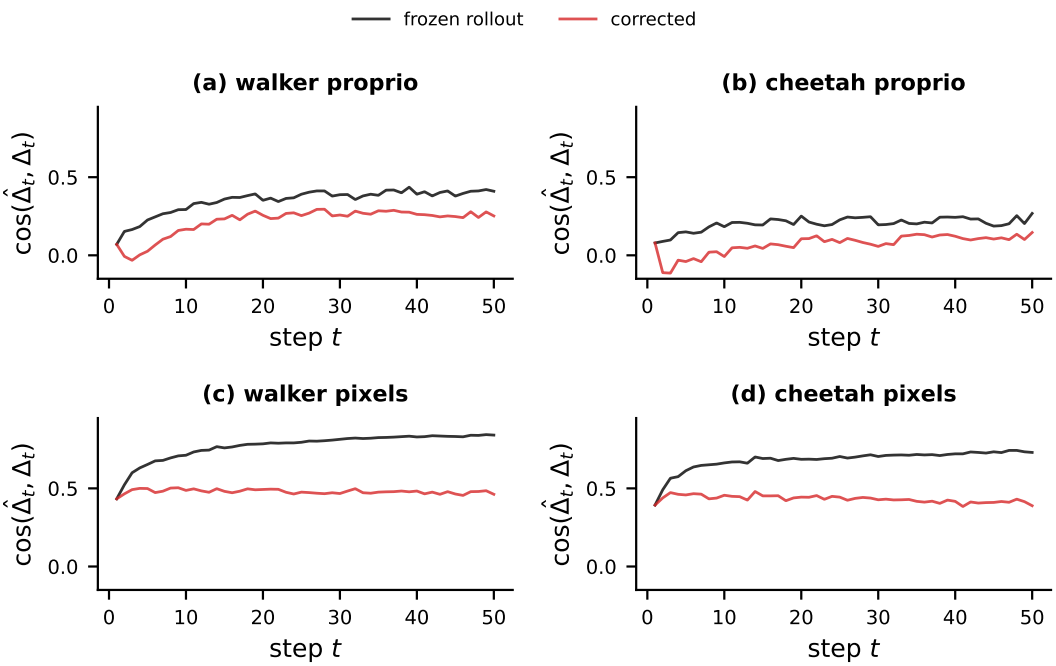



Figure 5: Information is spent once: probe alignment $\cos(\hat{\Delta}_t, \Delta_t)$ on frozen (gray) vs. corrected (red) rollouts. **(a)** Walker proprio. **(b)** Cheetah proprio. **(c)** Walker pixels. **(d)** Cheetah pixels. Alignment collapses within 2–4 steps after correction in all systems.

# 7 Discussion

## 7.1 Practical guidance: a two-dial audit

Our results compress into an audit that costs a few GPU-hours and no new method. Before buying test-time refinement, measure two dials. (a) **The budget**: fine-tune a copy of the WM on its own rollout contexts; $\sqrt{\text{FVU}_{\text{min}}}$ is the fraction of open-loop error any corrector could remove. (b) **The geometry**: $\cos(m, \Delta)$, whether the truth direction aligns with cheap return-to-manifold signals. High budget and high alignment (pixel regime) $\rightarrow$ refine; even zero-learning kNN-return recovers 22–31%, a probe corrector 29–46%. Low budget (near-manifold proprio) $\rightarrow$ do not; every method degrades fidelity. High budget but low alignment (starved proprio) $\rightarrow$ correction pays modestly ($-16\%$) only through a learned probe with small clamped steps. The audit also prices the train-time alternative: noise-injection (B2) consumes part of the same budget (Section 6 R4). One qualifier: pixel WMs are data-starved (val/train $\geq 14\times$), and the data axis shows the budget shrinking as data grows (FVU floor $0.31 \rightarrow 0.41$); with enough data the pixel regime migrates toward the proprio verdict.

## 7.2 Fidelity is not control

A corrector that improves imagination fidelity can still hurt the downstream planner; measured, not conjectured. In a paired model predictive path integral (MPPI) experiment on walker proprioception (exact latent reward, corr 1.0000; 16 episodes per arm) the fidelity-best corrector *significantly* degrades planning returns (probe $-$ raw $= -24.0$, 95% CI $[-46.6, -5.0]$) while improving one-step value calibration $\sim 5\times$ (Fig. 6). The mechanism: correction is legality-anchored, and a planner ranks futures precisely by how far they leave typical behavior; pulling off-distribution futures back toward the manifold is anti-exploratory.

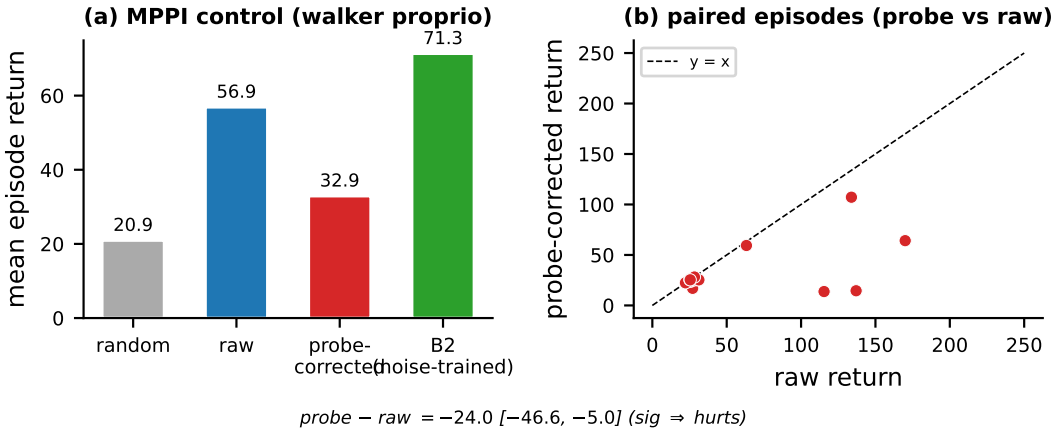



Figure 6: Fidelity $\neq$ control. **(a)** Mean MPPI returns (walker proprio). The fidelity-best corrector significantly hurts planning returns. **(b)** Paired episode returns (probe vs. raw); most points fall below $y = x$.

### 7.3 Related work

**Test-time verification (LLMs).** Process reward models (Lightman et al., 2023) and best-of-$N$ scaling (Cobbe et al., 2021; Snell et al., 2024) establish that verifiers convert test-time compute into quality. Our result is the continuous-control counterpart: a verifier can reach AUROC 0.94 and still carry zero information along the axis that matters.

**Energy-based model (EBM) guidance.** Energy composition (Du & Mordatch, 2019), classifier guidance (Dhariwal & Nichol, 2021), and safety projections (Chi et al., 2023) assume descending a learned constraint moves samples toward the desired set. We make that assumption measurable: $\cos(m, \Delta)$ is a property of the observation modality (0.11–0.73).

**Offline model-based RL.** MOPO/MOReL (Yu et al., 2020; Kidambi et al., 2020) *avoid* unreliable transitions; refinement *repairs* them. The budget law prices the repair route in advance. MBPO (Janner et al., 2019) truncates rollouts at train time; the budget is the test-time analogue.

**Train-time robustification.** Scheduled sampling / DAgger (Bengio et al., 2015; Ross et al., 2011), input-noise training, and hallucinated-replay (Talvitie, 2014; 2017) attack drift at train time. The $\Delta$-probe is DAgger-for-world-models, used here as a *measuring instrument* rather than a method.

**World models.** Our stack is the DINO-WM recipe (Zhou et al., 2024); Dreamer (Hafner et al., 2023), TD-MPC2 (Hansen et al., 2024), and IRIS/Genie (Micheli et al., 2023; Bruce et al., 2024) share the open-loop imagination our protocol measures.

### 7.4 Limitations

Three environments $\times$ two modalities (15 systems); humanoid is pixels-only, all environments are DeepMind Control (DMC) locomotion; no manipulation. Encoder attribution is addressed within the frozen-encoder family (ViT-S reproduces every sign) but all pixel latents are DINOv2 CLS vectors. All environments are deterministic; stochastic dynamics would add irreducible conditional noise (the law accommodates it; it lowers the budget). $\hat{B}$ is an estimate from above (Lemma 1): a stronger probe tightens it, and the one below-line point is adjudicated probe-limited.

## Acknowledgments

Anonymous submission.

### AI USE STATEMENT

We used generative AI tools (Claude, Anthropic) for retrieval and discovery, specifically for finding related work and identifying relevant citations during the literature survey. All retrieved references were manually verified for relevance and accuracy. The core contributions, experimental design, theoretical analysis, and writing are entirely our own. We take full responsibility for the final content of this work.

## REFERENCES

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *NeurIPS*, 2015.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Eoghan Hughes, Matthew Lai, Aadyot Mavalankar, Ryan Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *ICML*, 2024.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *RSS*, 2023.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat GANs on image synthesis. In *NeurIPS*, 2021.

Yilun Du and Igor Mordatch. Implicit generation and modeling with energy-based models. In *NeurIPS*, 2019.

Danijar Hafner, Jurgas Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Nick Hansen, Jie Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *ICLR*, 2024.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In *NeurIPS*, 2019.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. MoReL: Model-based offline reinforcement learning. In *NeurIPS*, 2020.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. In *ICLR*, 2023.

Stéphane Ross, Geoffrey Gordon, and J Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *AISTATS*, 2011.

Charlie Snell, Jaehoon Lee, Kelvin Zheng, and Ozan Sener. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Erin Talvitie. Model regularization for stable sample rollouts. In *UAI*, 2014.

Erin Talvitie. Self-correcting models for model-based reinforcement learning. In *AAAI*, 2017.

Tianhe Yu, Garrett Thomas, Luyu Xiong, Aravind Sudhakar, Rishabh Srinivasan, and Sergey Levine. MOPO: Model-based offline policy optimization. In *NeurIPS*, 2020.

Gaël Zhou et al. DINO-WM: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024.

Table 3: The law across 15 systems. Bound = $\sqrt{\min\text{-depth probe FVU}}$; achieved = best clamp/$\alpha$-schedule $\delta@50$ ratio vs. own raw. Gaps carry paired 95% CIs (10k bootstrap).

| **System** | $\hat{B}$ | Achieved | Gap |
|---|---|---|---|
| Walker proprio | 0.896 | 0.935 | $+0.039$ |
| Walker proprio (B2) | 0.941 | 0.968 | $+0.027$ |
| Cheetah proprio | 0.922 | 1.000 | $+0.078$ |
| Cheetah proprio (B2) | 0.972 | 0.974 | $+0.002$ |
| Walker pixels | 0.558 | 0.541 | $-0.017$ |
| Walker pixels (B2) | 0.659 | 0.698 | $+0.039$ |
| Cheetah pixels | 0.697 | 0.705 | $+0.008$ |
| Cheetah pixels (B2) | 0.789 | 0.801 | $+0.012$ |
| Walker proprio-100 | 0.880 | 0.841 | $-0.039$ |
| Walker pix-ViT-S | 0.632 | 0.626 | $-0.006$ |
| Walker pix-ViT-S (B2) | 0.752 | 0.740 | $-0.012$ |
| Walker proprio-300 | 0.880 | 0.844 | $-0.036$ |
| Walker pix-200 | 0.615 | 0.612 | $-0.003$ |
| Humanoid pixels | 0.639 | 0.624 | $-0.015$ |
| Humanoid pixels (B2) | 0.734 | 0.753 | $+0.019$ |

# A Detailed boundaries of the correction-budget law

**R0 — Corrector-induced distribution shift.** The proposition is stated at horizon $T$ through measurability: any closed-loop corrector's output $\tilde{z}_T$ is a (randomized) function of $\xi$ no matter how its intermediate corrections shift the rollout distribution; the proof never decomposes step-wise. Where the shift does matter is the mechanism of the above-bound gap: applying corrections consumes alignment (spent-once, Fig. 5).

**R1 — Information floor vs. learnable floor.** On deterministic, (near-)fully-observed proprioception $z_T$ is in principle $\xi$-measurable, so $\text{FVU}^*(\mathcal{G})$ can be $\approx 0$ and Prop. 1 alone is weak there. The paper's empirical content is the learnable floor: $\hat{B}$ measured with the WM's own architecture and data estimates what is statistically extractable. The probe-quality ladder (Table 4) makes the estimator-relativity explicit.

**R2 — Aggregate mismatch.** The theorem is RMS; measured $\delta$ is mean of norms. Since $\mathbb{E}\|X\| \leq (\mathbb{E}\|X\|^2)^{1/2}$, the mean-norm floor sits at or below the RMS floor. Confirmed: in RMS units all three mean-units below-line points become non-significant (walker100 $-4.0$pp $\to -1.3$pp; walker-pix $-1.7$pp $\to -1.0$pp; walker300 $-3.6$pp $\to -1.6$pp) and the law holds on 14/15 systems.

**R3 — Iterated correctors.** A closed-loop corrector composes the probe with the dynamics $T$ times; it remains in $\mathcal{C}$ but is not floored by the probe's own FVU. Hence achieved $< \hat{B}$ is possible without violation.

**R4 — Per-system statement.** Each WM defines its own $\hat{z}$, $\delta_{\text{raw}}$, and floor. B2's rollouts are less probe-predictable in 7/7 paired systems; input-noise training consumes part of the same budget at train time.

**R5 — What the law does not say.** It does not say refinement is useless (pixels: $-29\%$ to $-46\%$, bound saturated). It says the budget is fixed by open-loop predictability.

# B Run chronology and negative controls

## B.1 Chronology

The paper's claims were produced by twelve sequential go/no-go runs; each run's question was fixed before its numbers existed. Run-1: verifier fix (AUROC 0.06 $\to$ 0.88). Run-2: data volume (120 eps inverts). Run-3: directional verifier (cos 0.03 $\to$ 0.14, still fails). Run-4: particle selection (legality improves, truth degrades). Run-5: $\Delta$-probe (FVU floors, two channels). Run-6: replication battery

Table 4: Probe-quality ladder on walker100 (pre-registered). The bound tightens monotonically with probe quality; achieved tracks it ≈1:1.

| **Probe quality** | $\hat{B}$ | Achieved |
|---|---|---|
| Weak (3k steps, 4k window) | 0.915 | 0.888 |
| Medium | 0.870 | 0.836 |
| Canonical (6k, 16k) | 0.880 | 0.841 |
| Strong (12k, 32k) | 0.870 | 0.836 |
| Strongest (24k, 64k) | 0.860 | 0.811 |

(proprio + pixels). Run-7: cheetah pixels (pixel regime replicates). Run-8: probe-quality ladder. Run-9: ViT-S encoder attribution. Run-10: MPPI control (fidelity $\neq$ control). Run-11: data axis (proprio-300, pix-200). Run-12: humanoid pixels (third pixel env).

### B.2 Negative controls

- **Double-counting control** (`sim_budget.py`): per-step-skim simulator re-draws alignment every step, over-predicts cash-in 3×.
- **Seed audit**: training seeds $\{0, 1, 2\}$ leave geometry dial within SD 0.011/0.005 (proprio/pixels); $\hat{B}$/achieved within 0.6/1.4pp.
- **Pre-registration**: all ladders and replication batteries carry predictions in script headers, written before numbers existed.
- **Oracle sanity**: unclamped O2 reaches ratio 0.098 on walker, far below every bound, as required for an oracle outside $\mathcal{C}$.

## C Reproduction details

- **Environment**: single H100; Python 3.10, PyTorch $\geq$ 2.4; dm_control for proprio; V-D4RL for pixels; frozen DINOv2 staged into repo-local TORCH_HOME.
- **Canonical protocol**: context $H = 16$; horizon $T = 50$ (drift-shape checks at $T = 200$); probe depths $\{0, 1, 2, 4, 8, 16, 24, 32, 39\}$; proprio $n = 1024$, pixels $n = 256$.
- **Artifacts**: filenames encode configuration; every number in the paper appears in a JSON under `runs/eval/`.