CALIBRATED UNCERTAINTY FOR LONG-HORIZON ROLLOUTS IN LATENT WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Latent world models can imagine long rollouts, but planners need to know when their predictions become unreliable. We show that common uncertainty measures, including ensemble disagreement, latent divergence, and predictive variance, are Lipschitz functions of compact latent states and therefore have horizon-independent ceilings, while rollout error need not. Across two world-model families, ensemble disagreement grows more slowly than reward error and becomes anti-correlated with it on several tasks. We address this mismatch without retraining or ensembles. A linear probe maps a frozen latent state to rollout error and is calibrated by split conformal prediction at each horizon. The resulting intervals achieve finite-sample coverage and track per-rollout error better than standard conformal baselines on nearly every DreamerV3 and TD-MPC2 task. Where error compounds, they are substantially sharper than ensemble-scaled intervals, while matching them when error remains flat. A single-checkpoint probe matches an ensemble-trained probe at $1/K$ of the training cost with a K-member ensemble. For trusted-horizon decisions, our intervals accept as many imagined steps as ensemble disagreement while exceeding the error tolerance less often. Coverage can degrade under substantial behavior-policy shifts, requiring recalibration.

1 INTRODUCTION

World models that learn latent dynamics now set the standard for sample efficiency in model-based reinforcement learning. DreamerV3 (Hafner et al., 2024), DreamerV2 (Hafner et al., 2021), MuZero (Schrittwieser et al., 2020), TD-MPC2 (Hansen et al., 2024) and the JEPA family (Assran et al., 2023) all learn to imagine rollouts in a compact latent space. Several uses of those rollouts need an uncertainty estimate that can be trusted at long horizons, among them risk-sensitive planning (Webster & Flach, 2021), optimistic exploration (Seyde et al., 2021) and safe policy improvement under distribution shift. The standard estimate is ensemble disagreement (Lakshminarayanan et al., 2017). One trains several models, rolls them forward from the same state, and reads their spread as epistemic uncertainty. In observation space, this works, and both PETS (Chua et al., 2018) and the more recent RWM-U (Li et al., 2026) use it for control over long autoregressive rollouts. However, the picture is different in latent world models.

Berger et al. (2026) documented this failure. In Dreamer's RSSM, ensemble disagreement over predicted latents does not reliably reflect compounding model error over long rollouts. They attribute this to an attractor effect: rollouts are pulled toward well-supported latent regions, where uncertainty shrinks while true divergence grows. They call it a structural limitation that better uncertainty estimation alone cannot resolve. Their evidence is qualitative, over four DMC tasks, and they ask for a "formal characterization of the attractor mechanism." Separately, Aadithiyan (2026) found a linear probe on the RSSM recurrent state that detects model confusion at AUROC 0.72 where ensemble disagreement is at chance (0.55). Down-weighting imagined value estimates by that probe made them worse. A better latent signal, fed to the planner, did not produce better decisions.

Contributions. We address both the theoretical gap and the practical impasse. We prove that the RSSM's deterministic state is confined to $[-1, 1]^D$ by its sigmoid gating and tanh candidates, and that its stochastic state lies on a finite simplex product (Propositions 1, 2). Latent divergence between any two rollouts is therefore bounded by a constant independent of the horizon (Corollary 3). Any uncertainty statistic that is Lipschitz in latent discrepancy inherits this ceiling; physical error has none (Theorem 4). This applies to ensemble disagreement, latent KL and predictive variance.

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.

The bound limits magnitude rather than ordering, since correlation is scale-invariant and a bounded statistic can still rank rollouts by an unbounded error. Which bounded statistics actually track error is decided by supervision rather than by the bound, and the rest of the paper turns on that distinction. We then measure what the bound costs. On most DreamerV3 tasks where error compounds, disagreement grows more slowly than reward error and stops tracking it; it stays informative only on the tasks where it keeps pace. We report that as a measured regularity rather than a mechanism. We tested two candidate mechanisms, contraction of the latent update and collapse of imagined reward diversity, and falsified both.

The repair needs no retraining, no change of architecture and no ensemble, which is what makes it cheap. The ensemble baselines train K world models where the probe reads one, in exchange for about eight rollouts whose true reward was recorded. We fit a linear probe of the frozen latent to measured error, then calibrate it by split conformal prediction (Lei et al., 2018) at each horizon. The probe supplies the ordering a bounded latent can still provide, which rollouts are harder, and the conformal quantile supplies the magnitude it cannot. We evaluate on two world-model families. DreamerV3 covers 40 tasks (11 DMC proprioceptive, 4 DMC visual and 25 Atari 100k games) and TD-MPC2 covers 19 DMC tasks, with Crafter as a further row. Coverage stays between 0.89 and 0.91 against the 90% target on every task. Interval width tracks per-rollout error better than every conformal baseline we test, on nearly every task. Measured by interval score, the gain over a stabilized ensemble-scaled interval is concentrated where error compounds. On those 17 DreamerV3 tasks the probe wins 14, at a median $0.66\times$ the ensemble's score, and on flat-error tasks the two are equivalent. We also report where the approach does not help.

In summary, model-internal uncertainty fails at long horizons because of how the compact latent is read, not because the latent lacks the information. A linear probe of a frozen model, calibrated on held-out episodes, recovers the missing magnitude, and it is more adaptive than the scale of a K-member ensemble on 27 of 40 DreamerV3 tasks and 15 of 19 TD-MPC2 tasks. A probe fitted on a single checkpoint matches one fitted on the ensemble mean, so the method needs one trained model where the baseline needs K. As a decision signal, the intervals answer one question directly, which is how many imagined steps deserve trust, and they accept as many steps as the ensemble scale while violating the error tolerance less often. Predictions and pass bars were fixed before the runs that tested them, and negative results are reported alongside positive ones.

2 Related Work

Uncertainty in model-based RL. Ensemble disagreement (Lakshminarayanan et al., 2017) is the standard epistemic estimator. PETS (Chua et al., 2018) propagates it through observation-space dynamics for planning, and MBPO (Janner et al., 2019) limits compounding error by keeping rollouts short. LOVE (Seyde et al., 2021) extends ensembles to latent models but consumes the uncertainty optimistically, for exploration. RWM-U (Li et al., 2026) propagates uncertainty over long autoregressive rollouts for risk-averse control, from raw observations rather than a compact latent, which by Theorem 4 is the setting where it escapes the ceiling we identify. Webster & Flach (2021) penalize planning by ensemble variance, DreamerV3-XP (Bierling et al., 2025) uses disagreement for exploration, and WIMLE (Aghabozorgi et al., 2026) trains uncertainty-aware world models by importance-weighted likelihood. None of them ask whether the signal is still informative at the horizons where compounding error dominates.

Latent uncertainty and its limits. Berger et al. (2026) are the closest antecedent. They document that disagreement in Dreamer's RSSM does not reflect compounding error over long rollouts, attribute it to an attractor effect, and call it a limitation that "cannot be resolved through improved uncertainty estimation alone." We supply part of the formal characterization they ask for, as a compactness bound, and measure where it binds across 60 tasks. We could not establish the mechanism behind it. Aadithiyan (2026) found a latent probe that detects model confusion yet hurt value estimates when used to down-weight them; our results suggest why, since the probe gives ordering but not magnitude. Yu et al. (2026a) attack the same limitation from the other side, reshaping the latent so that it stays sensitive to control inputs where the data supports it. We leave the model untouched and calibrate a readout of it instead.

Decision-aware model learning. A separate line argues that model objectives should reflect how the model is used. Value-aware model learning and its calibrated variant (Voelcker et al., 2025) weight model error by its effect on the value estimate, the value equivalence principle (Grimm et al.,

2020) formalizes being accurate only where it changes decisions, λ-models (Voelcker et al., 2024) implement decision-aware RL with latent models, and Voelcker et al. (2023) weight the model loss by value sensitivity. Our work is complementary, since we change neither the objective nor the model and instead calibrate a post-hoc readout of a frozen one.

Conformal prediction and world-model evaluation. Split conformal prediction (Lei et al., 2018) gives distribution-free intervals under exchangeability, conformalized quantile regression (Romano et al., 2019) extends it to heteroscedastic ones, and CF-RNN (Stankevičiūtė et al., 2021) covers a whole time series by a Bonferroni correction, which we evaluate in Appendix G. On the evaluation side, Yu et al. (2026b) propose a ladder of world-model evaluations that includes uncertainty calibration, and Schneider et al. (2026) argue for decision-aligned metrics in supervised prediction. Our evaluation sits on that calibration rung and extends it to long-horizon rollouts.

3 Background

Recurrent state-space models. DreamerV3 (Hafner et al., 2024) learns dynamics in a latent state-space model that pairs a deterministic recurrent state $h_t \in \mathbb{R}^D$ with a categorical stochastic state $z_t \in \mathbb{R}^{S\times C}$, for S groups of C classes. The deterministic state follows a gated update

$$h_t = u_t \odot c_t + (1-u_t)\odot h_{t-1}, \qquad u_t = \sigma(x_u - 1), \qquad c_t = \tanh\big(\sigma(x_r)\odot x_c\big), \tag{1}$$

where σ is the elementwise sigmoid and x_u, x_r, x_c are linear projections of the previous state and action. The stochastic state is drawn from a categorical prior $\pi(h_t)$ and is one-hot per group, and during imagination only that prior is used, so a rollout never sees an observation. The configuration we evaluate uses $D = 512$, $S = 32$ and $C = 4$, with a reward head trained by two-hot encoding on a fixed 255-bin support (Appendix E).

Split conformal prediction. Split conformal prediction (Lei et al., 2018; Angelopoulos & Bates, 2021) turns a nonconformity score $s(x, y)$ into intervals with finite-sample marginal coverage. Given n calibration pairs and a level α, it takes the $\lceil (n+1)(1-\alpha)\rceil/n$ quantile $\hat{q}$ of the calibration scores and returns $\hat{C}(x) = \{y : s(x, y) \le \hat{q}\}$. Under exchangeability, $P(y_{\text{test}} \in \hat{C}(x_{\text{test}})) \ge 1-\alpha$ for any n and any distribution. Such intervals are marginally but not conditionally valid: one global quantile gives easy and hard inputs the same width. Adaptive variants replace it with an input-dependent quantile at the cost of weaker guarantees (Lei & Wasserman, 2014; Romano et al., 2019; Vovk et al., 2005). We calibrate separately at each horizon, since one global quantile would over-cover early and under-cover late.

4 Why Latent Uncertainty Cannot Track Long-Horizon Error

The RSSM latent space is compact as a consequence of the normalization that makes the model trainable, and that compactness limits any model-internal uncertainty statistic. Both states are bounded (Propositions 1, 2), so latent divergence is bounded by a horizon-free constant (Corollary 3) and any Lipschitz statistic of it runs out of range against an unbounded physical error (Theorem 4). Split-conformal validity survives any rescaling of the score (Proposition 5), so a supervised readout can restore the magnitude without losing the guarantee. Proofs appear in Appendix A.

Proposition 1. *For the RSSM update in equation 1, $h_t \in [-1, 1]^D$, $\forall t \ge 0$, every input sequence, and every parameter setting.*

Algorithm 1 Probe-scaled conformal intervals. Nothing in the world model is retrained.

Require: frozen world model, logged episodes, horizon H, level α

1: Split episodes three ways: TRAIN, CAL, TEST.
2: From each real state, imagine H steps and execute the same actions to obtain the measured error $y_h = |\hat{R}_h - R_h|$ and the frozen latent x_h.
3: Fit $\hat{w}$ by ridge regression of y_h on the standardized x_h over TRAIN only (Eq. 2).
4: Set the scale $s(x) = \max(\hat{w}^\top x, 0) + \beta\bar{p}$, with β chosen on TRAIN to minimize mean width.
5: **for** $h = 1$ **to** H **do**
6: $\quad \hat{q}_h \leftarrow \lceil (n+1)(1-\alpha)\rceil/n$ quantile of $\{\, y_{h,i}/s(x_i) : i \in \text{CAL} \,\}$
7: **end for**
8: **return** the interval $[\,0,\ s(x)\,\hat{q}_h\,]$ for a TEST rollout at horizon h

Proposition 2. *The stochastic state z_t is a concatenation of S one-hot vectors over C classes, so $\|z_t\|_2 = \sqrt{S}$, $\forall t$, and for any two rollouts $\|z_t - z'_t\|_2 \leq \sqrt{2S}$.*

Neither is a property of training. No learned weight can violate them, because $\tanh$ and σ bound the two ingredients of the convex combination, and the straight-through estimator (Bengio et al., 2013) keeps z_t one-hot in the forward pass. Together they bound how far two rollouts can drift apart in latent space.

Corollary 3. *For any two latent trajectories, $\left\|(h_t, z_t) - (h'_t, z'_t)\right\|_2 \leq \sqrt{4D + 2S}$ for all t. At the configuration we evaluate ($D = 512$, $S = 32$) this is $\sqrt{2112} \approx 45.96$, independent of the rollout horizon.*

Theorem 4. *Let U_t be any uncertainty statistic that is L-Lipschitz in latent-state discrepancy, so that $|U_t - U'_t| \leq L \|(h_t, z_t) - (h'_t, z'_t)\|_2$. Ensemble disagreement over predicted latents, KL divergence between prior distributions, and latent predictive variance are all of this form. By Corollary 3, $U_t \leq L\sqrt{4D + 2S} =: U_{\max}$ for all t. Let $\delta_t = \|s_t - \hat{s}_t\|$ be the physical discrepancy between true and predicted states, which is unbounded whenever the true dynamics are not globally stable. Beyond the horizon at which $\delta_t > U_{\max}/L$, U_t cannot order trajectories by physical error even in principle, because the estimator has run out of range.*

Decoding does not escape this. The decoder, the reward head and the value head are continuous maps of a compact set, so their outputs are bounded as well, and disagreement read in observation or reward space inherits the same ceiling (Corollary 6). Reward and value disagreement are covered too, since `symexp_twohot` places predicted returns on a fixed 255-bin support.

Proposition 5 (Validity is scale-invariant)**.** *Fix a horizon h and a level α, and let $s(x) > 0$ be any measurable scale that is fitted without reference to the calibration or test episodes. Write $r_i = y_i/s(x_i)$ for the ratio score. If the calibration scores $\{r_i\}_{i \in \text{cal}}$ and the test score r are exchangeable conditional on s, then the interval $\hat{C}(x) = [0,\, s(x)\,\hat{q}_h]$, with $\hat{q}_h$ the $\lceil (n+1)(1-\alpha) \rceil / n$ quantile of the calibration ratios, satisfies $P\big(y \in \hat{C}(x)\big) \geq 1 - \alpha$.*

Ensemble disagreement, the prior KL and the predictive variance of any bounded head are all Lipschitz in the same latent discrepancy, and Corollary 6 extends the argument to statistics read after decoding. TD-MPC2's SimNorm latent, a product of simplices, is compact for the same reason (Appendix F.3), so the ceiling is a property of the design pattern rather than of one architecture. Appendix D gives the derivations. Boundedness does not make a latent statistic uninformative. Correlation is scale-invariant, and our linear probe reaches late-horizon correlations as high as $+0.990$. What boundedness prevents is recovering the scale of an unbounded error, which Proposition 5 shows can be restored by calibration on held-out episodes. Thus, the latent provides the ordering and the conformal quantile provides the missing magnitude, which is the design principle behind Section 5.

Practicality of the bound. Theorem 4 is asymptotic; the form that applies to bounded benchmarks is about growth rates. Latent divergence reaches its effective ceiling almost immediately, and we measure it at 6.65, flat to three digits from $h = 1$ to $h = 50$, while physical error keeps growing, and a constant statistic carries no information about a changing error.

5 Probe-Scaled Conformal Intervals

Overview. Given a frozen world model, we collect imagined rollouts from real states, record the ground-truth reward error at each horizon, and fit a linear probe of the frozen latent to predict it. That probe output becomes the multiplicative scale of a per-horizon conformal interval (Figure 1). From each real state in the evaluation set we imagine $H = 50$ steps with the frozen model's prior and record the predicted reward, then execute the same actions in the real environment to obtain the true reward. The target is $|\hat{R}_h - R_h|$, the absolute error in cumulative predicted reward at horizon h. We read the frozen imagined latent at each horizon and fit ridge regression against measured error on train episodes only. For each horizon we then take the $\lceil (n+1)(1-\alpha) \rceil / n$ quantile of $\{|\hat{R}_{h,i} - R_{h,i}| / \text{probe}(x_i)\}$ over calibration episodes, and the interval for a test rollout is $\text{probe}(x) \cdot \hat{q}_h$.

The three-way episode split into train, calibration and test keeps the probe from ever seeing the episodes it is calibrated or scored on. Rollouts within an episode are correlated, so splitting by

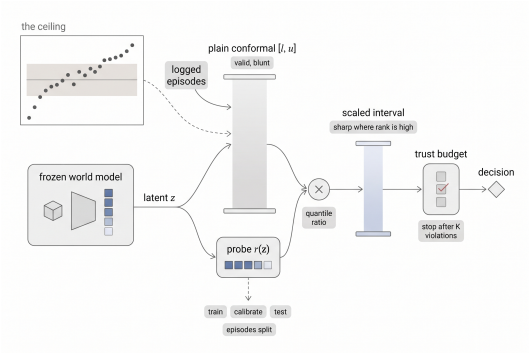

Figure 1: The ceiling and the repair. Inset: physical error grows with the imagined horizon while any statistic of the compact latent stays within a horizon-free band. Main panel: a linear probe of the frozen latent gives the per-rollout scale $s(x)$, per-horizon split conformal gives the quantile $\hat{q}_h$, and the interval is $[0,\ s(x)\hat{q}_h]$; the trust budget keeps the longest prefix whose width stays under the tolerance. Episodes split three ways: train fits the probe, calibration fits $\hat{q}_h$, test scores.

episode rather than by rollout is what makes the exchangeability assumption behind conformal coverage reasonable.

The probe. The probe is a linear readout of the frozen imagined latent at horizon h (for DreamerV3 the concatenation of the deterministic and stochastic states). In the main tables, where an ensemble of checkpoints is available for the baselines, the latent and the predicted reward are averaged over members, and the target is the error of the member-mean reward. The method does not need the ensemble. A probe on one checkpoint's own latent, predicting that checkpoint's own error, reaches the same late-horizon adaptivity (mean 0.753 vs. 0.752 over the 29 tasks with per-member dumps) with coverage 0.894-0.910. Features are standardized across all rollouts before fitting. We use ridge regression with an unpenalized intercept and a penalty scaled by the number of training rows, which makes the regularization strength dimensionless.

$$\hat{w} = \arg\min_{w} \sum_{i \in \text{train}} \left(y_i - [x_i; 1]^\top w\right)^2 + \lambda n \, \|w_{1:d}\|_2^2, \tag{2}$$

where $x_i \in \mathbb{R}^D$ is the standardized latent feature, $y_i = |\hat{R}_{h,i} - R_{h,i}|$ is the measured reward error, and $w_{1:d}$ excludes the intercept from the penalty. The penalty λ is selected per task by the calibration rule of the hyperparameter sweep, which never uses test episodes.

The probe output is floored at zero and stabilized by adding $\beta \cdot \bar{p}$. Here $\bar{p}$ is the mean probe output on train episodes, and $\beta \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1.0\}$ is selected on those same episodes to minimize mean interval width. Without this floor the conformal quantile diverges when the probe predicts near-zero error for easy rollouts.

Feature variants. The main results use the latent feature set, a pure readout of the frozen latent, which is what the theory speaks to. Adding the normalized horizon index and ensemble disagreement gives a richer set that tests whether the probe can absorb the unsupervised signal. On reacher_easy, the one task where disagreement clearly beats the latent-only probe, it takes late adaptivity from $+0.569$ to $+0.900$ (Appendix F.5). The comparison is therefore not probe versus disagreement, but a supervised readout against the best signal available to it.

Conformal calibration. For horizon h and scale $s(x)$ the interval is

$$\hat{C}(x,h) = \left[0,\ s(x)\cdot\hat{q}_h\right], \qquad \hat{q}_h = \text{Quantile}_{(n+1)(1-\alpha)/n}\left(\left\{\frac{|\hat{R}_{h,i} - R_{h,i}|}{s(x_i)} : i \in \text{calibration}\right\}\right), \tag{3}$$

with $s(x) = \text{probe}(x) + \beta\bar{p}$ for our method, $s(x) = \max(\text{disagreement}(x), \epsilon)$ for ensemble-scaled conformal, $s(x) = 1$ for plain conformal and $s(x) = \text{kNN}(x)$ for the locally weighted variant. Calibrating separately at each horizon matters, because the relationship between probe output and error changes with h.

Baselines. We compare against five conformal variants, namely plain conformal ($s(x) = 1$), ensemble-scaled, Mondrian (binned on disagreement), k-NN local conformal (Lei & Wasserman, 2014) and CQR (Romano et al., 2019), and against an oracle probe fitted in-sample on the test

episodes, which bounds what any linear readout of these features can do. Where ensemble disagreement collapses toward zero, its interval blows up, so we report a stabilized ensemble scale that uses the same train-selected floor as the probe; that is the stronger comparator and the one we report against throughout. We also compare against published uncertainty signals used as the conformal scale. Appendix I specifies each.

6 EXPERIMENTS

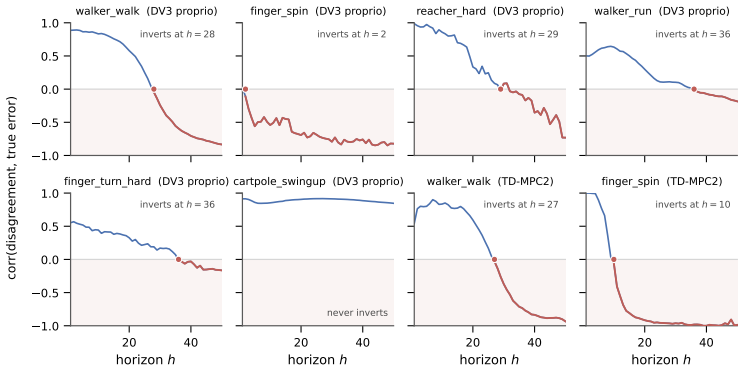

Figure 2: The disagreement–error correlation by horizon, for the five DreamerV3 tasks whose error grows and whose late correlation is non-positive, three contrast rows, and two TD-MPC2 panels. The dot marks the crossing; the curve turns red and the half-plane is shaded where disagreement is anti-correlated with the error it should track.

Setup. We evaluate 40 DreamerV3 tasks, made up of 11 DMC (Tassa et al., 2018) proprioceptive tasks, 4 DMC pixel tasks and 25 Atari 100k (Kaiser et al., 2020) games. Freeway is excluded because DreamerV3 scores zero on it at this budget, leaving every correlation undefined. To test whether the findings are specific to DreamerV3 we repeat everything on TD-MPC2 (Hansen et al., 2024), whose latent is SimNorm-normalized rather than recurrent, over 19 DMC tasks, and add Crafter (Hafner, 2022) as a further row. Models are trained for 1M environment steps on DMC and 100k on Atari, with $K = 5$ or $K = 3$ seeds per task as ensemble members, and every checkpoint was verified to have reached its budget. For each task we collect 40 evaluation episodes (94 for Crafter, 60 to 80 for TD-MPC2), imagine $H = 50$ steps from each real state, and score against the true reward obtained by executing the same actions. Intervals are calibrated at the 90% level over 120 random three-way episode splits, with late-horizon statistics over $h \in [25, 50]$. Three auxiliary analyses use 60 splits and are named in Appendix E. We report three quantities throughout. Adaptivity is the correlation across rollouts between interval width and true reward error, averaged over late horizons. The hard third is the third of rollouts with the largest mean error, on which we report coverage separately, since marginal coverage can be held by over-covering easy rollouts. The interval score (Gneiting & Raftery, 2007) is a proper score that charges width and penalizes non-coverage at $1/\alpha$, so it cannot be improved by widening or by narrowing alone. A task's error compounds if it grows significantly with the horizon under a bootstrap over episodes. Counts are tested with a two-sided exact sign test over tasks and ratios carry a bootstrap interval. Predictions and pass bars were fixed before the runs that tested them (Appendix E).

Disagreement inversion. The correlation between ensemble disagreement and true reward error decays with the rollout horizon, and on some tasks it changes sign. On 7 of the 40 DreamerV3 tasks, and on 7 of the 19 TD-MPC2 tasks, the correlation is positive at early horizons and non-positive at late ones, so the signal ends up anti-correlated with the error it should track (Figure 2). Walker_walk falls from $+0.891$ early to -0.791 late, and finger_spin sits at chance at $h = 1$, crosses at $h = 2$, and worsens to -0.893. On 23 of the 40 DreamerV3 tasks the error does not grow significantly under a bootstrap over episodes, so there is nothing to track and the question cannot be answered. Those are concentrated in Atari, where 21 of 25 games have flat error, and that flatness is a property of the games rather than of agent strength or horizon. A pong agent trained for $10\times$ the budget does not change the picture. Extending to $H = 150$ gives significant error growth in only 3 of the 21 flat games, 2 of the 18 not already tested at that horizon, against a pre-registered prediction of 3 to

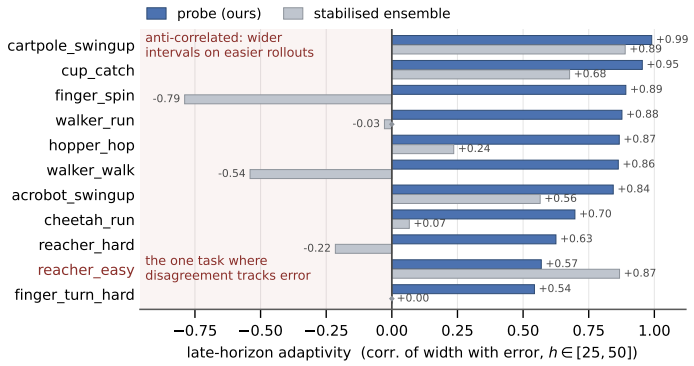

Figure 3: Late-horizon adaptivity (the correlation between interval width and true reward error) for the 11 DMC proprio tasks, the stabilized ensemble beside the probe. Shading marks the anti-correlated half-plane, where a method gives its wider intervals to the easier rollouts.

9. Counting inversions among only the tasks whose error grows, so that the question is answerable, leaves 5 on DreamerV3 and the same 7 on TD-MPC2. Table 4 in Appendix F lists them. Every unsupervised statistic read from the compact latent inherits the same ceiling (Theorem 4), though the ceiling alone does not predict which tasks invert.

Validity, and where it is hard to hold. Table 1 reports coverage, adaptivity, and interval score for the probe and every conformal baseline. Coverage stays between 0.901 and 0.914 against the 90% target on all 40 DreamerV3 tasks, so the per-horizon guarantee holds (Appendix F.4). The harder question is whether an interval still covers the rollouts the model finds hardest, and there the scales separate. On the hardest third of rollouts, the probe covers 0.766 against plain conformal's 0.724 on DreamerV3, and 0.785 against 0.734 on TD-MPC2, with the stabilized ensemble in between. Adaptivity is positive on every task, from $+0.066$ on demon_attack to $+0.990$ on cartpole_swingup (Figure 3). It is more adaptive than every conformal baseline on at least 27 of 40 DreamerV3 tasks and 15 of 19 TD-MPC2 tasks, at coverage between 0.893 and 0.907 (Table 1). Mondrian achieves

Table 1: Per family and scale: marginal coverage against the 90% target, coverage on the hardest third of rollouts, late-horizon adaptivity, and median interval score relative to plain conformal. The ensemble row is the stabilized scale. Paired counts: tasks on which the probe is more adaptive (left) and scores lower (right).

Family	Scale	Cov.	Hard third	Late adapt.	IS/conf	Probe better adapt.	Probe better IS
DreamerV3	Probe (ours)	0.906	0.766	+0.552	0.96		
	Probe + disagreement	0.906	0.768	+0.587	0.94		
	Ensemble-scaled	0.907	0.728	+0.230	0.99	27/40	22/40
	k-NN local	0.906	0.753	+0.414	0.98	38/40	33/40
	Mondrian	0.921	0.826	+0.176	0.98	40/40	29/40
	Plain conformal	0.908	0.724	+0.000	1.00	40/40	37/40
TD-MPC2	Probe (ours)	0.902	0.785	+0.745	0.90		
	Probe + disagreement	0.902	0.784	+0.803	0.90		
	Ensemble-scaled	0.905	0.740	+0.262	0.99	15/19	11/19
	k-NN local	0.903	0.772	+0.605	0.99	18/19	11/19
	Mondrian	0.922	0.804	+0.345	0.95	18/19	9/19
	Plain conformal	0.912	0.734	+0.000	1.00	19/19	15/19

the best hard-third coverage of any scale we test, and Table 10 shows what that costs. A median width of $1.10\times$ plain conformal conceals a mean of $44\times$, nine of forty rows beyond $10\times$, and a worst row of $1012\times$ ($107\times$ on TD-MPC2), while the probe's worst row is $1.4\times$. Among the scales that hold target coverage with bounded width, the probe covers the hard third best. Its interval score is lower than plain conformal's on 37 of 40 DreamerV3 tasks (15 of 19 on TD-MPC2).

Quantile regression is the informative comparison among these, because it is the one baseline that also fits a supervised model of the residual rather than reading a statistic from the model. It holds coverage on every task, so it is not failing on validity, and it still pays a median $2.33\times$ the plain-

conformal width for its adaptivity where the probe pays $0.85\times$. Pooled over both families, the count is 57 of 58 (two-sided exact sign test $p = 4 \times 10^{-16}$, Table 5), so a valid adaptive baseline is losing on ordering, not on validity. Fitting the residual is not by itself enough; what the latent supports is an ordering, and the conformal quantile is what turns that ordering into a width.

Against a stabilized ensemble scale. Against the stabilized ensemble scale, which is the stronger comparator, the probe is the more adaptive scale on 27 of the 40 DreamerV3 tasks (two-sided sign test $p = 0.039$) and on 15 of 19 TD-MPC2 tasks ($p = 0.019$), and it does so from a single checkpoint where the ensemble needs K. Sharpness separates by regime rather than uniformly. Over all 40 DreamerV3 tasks the probe's interval score is lower on 22 ($p = 0.64$). On the 17 where error compounds, it wins 14, at a median ratio of 0.66 with a 95% bootstrap interval of 0.53 to 0.86 ($p = 0.013$); on the 23 flat-error tasks it wins 8 at a median 1.01, and the two are equivalent. On TD-MPC2 it wins 11 of 19 at a median 0.88 (Figure 5). That split is what the bound predicts. Where error does not grow, there is nothing for an adaptive scale to track. Aggregated by horizon over the 36 compounding settings, the probe's median width is below plain conformal's in every horizon window, dipping to $0.747\times$; against the stabilized ensemble it falls to $0.338\times$ at $h \geq 41$ (Figure 4).

Robustness. *Ensemble size.* Rebuilding both arms on K-member subsets of the per-member checkpoints, all three pre-registered bars pass; at $K{=}1$ the ensemble scale degenerates to plain conformal, so the two arms start from the same point (Appendix F.7). *Retraining.* On three tasks we retrained three disjoint 3-seed ensembles each; absolute adaptivity is seed-sensitive (split resampling understates the spread by up to $4.3\times$), but the disagreement-augmented probe still beats the stabilized ensemble on interval score in 8 of the 9 groups, against a pre-registered bar of 7. *One checkpoint.* Fitted on a single checkpoint's own latent against that checkpoint's own error, with no ensemble anywhere, the probe holds coverage on 29 of 29 rows and covers the hard third at 0.830 against plain conformal's 0.747 (Appendix F.7). Its late adaptivity matches a probe fitted on the ensemble mean, 0.753 against 0.752. The comparisons above therefore hold at $1/K$ the training cost, and the probe's advantage grows as K falls, with a median ratio to the stabilized ensemble of 0.587 at $K{=}2$ against 0.778 at the full ensemble. Where an ensemble has been trained anyway, its disagreement can be an input to the probe rather than a competing scale, which is more adaptive than the stabilized ensemble on 35 of 40 DreamerV3 tasks against 27 for the latent-only probe (Appendix F.5).

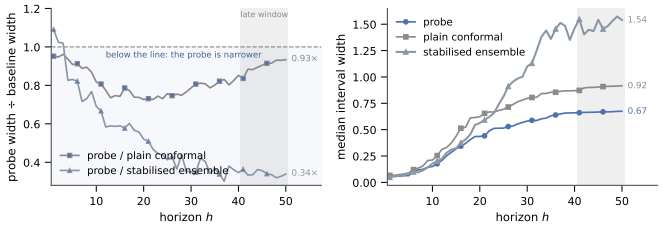

Figure 4: Interval width by horizon on the 36 live settings where error compounds: the median width of plain conformal, the stabilized ensemble and the probe (right), and the probe's width relative to each (left). Shading marks the late window $h \in [41, 50]$.

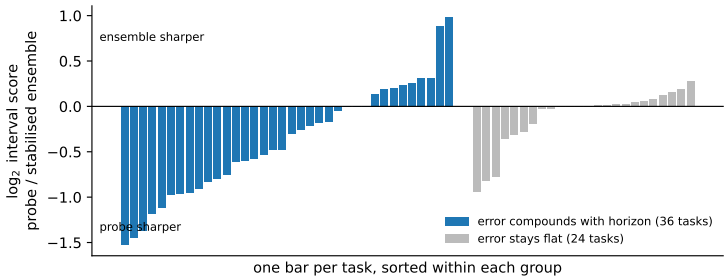

Figure 5: Interval score of the probe relative to the stabilized ensemble scale, one bar per task, sorted within each group; the scale is logarithmic and the dashed line is each group's median.

Data needs and published signals. The probe is fitted against measured error and needs little of it, since 46 of 60 tasks are within 10% of full adaptivity by 8 training episodes. Coverage stays within

± 0.02 of target at every budget down to a single episode that is the split-conformal guarantee, not the fit. Scored against four unsupervised signals, the probe is best on 33 of 60 tasks. Against the two single-model signals that need no ensemble, the latent-transition disagreement of Berger et al. (2026) and a last-layer Laplace approximation of the reward head, it has the lower interval score and the higher adaptivity on all 11 proprioceptive tasks (Appendices F.1 and F.6).

Deciding how far to trust a rollout. An interval that describes a rollout well need not tell a planner what to do, and in our experiments it does not. Appendix G reports every way we tried of turning these intervals into a control signal, and all are null. What they do support is a budget. For each test rollout we keep the longest prefix whose interval width stays below a tolerance ϵ, set per split from plain-conformal widths on train rollouts, which is independent of the scale under test (Table 2).

Table 2: Steps trusted at the median tolerance and the rate at which an accepted step breaks it. A superscript marks the rows on which the violation rate is defined, since a method that accepts no step has none.

	Probe (ours)		Stabilised ensemble		Plain conformal	
Family	steps	violation	steps	violation	steps	violation
DreamerV3 (40 tasks)	12.6	0.027^{38}	13.0	0.048^{38}	8.8	0.067^{34}
TD-MPC2 (19 tasks)	27.5	0.019	25.9	0.023	24.1	0.034

Against plain conformal the probe trusts a longer prefix at no more than 0.02 higher violation. That holds on 28 of the 33 DreamerV3 tasks where both accept a step, at a median $1.42\times$ (two-sided sign test $p = 10^{-4}$), and on 14 of 19 TD-MPC2 tasks at a median $1.30\times$ ($p = 0.064$), where the effect size rather than the count carries the family. Against the stabilized ensemble it does not trust more steps, and our pre-registered criterion for that comparison failed. The probe wins on 26 of the 53 tasks where all three scales accept a step, against a bar of 35. What it does is accept the same number of steps with fewer violations, 0.027 against 0.048 on DreamerV3. On the one flat Crafter task, which the table omits, the ensemble accepts more steps at the same violation rate.

Limitations. The probe is linear, yet reaches a median 87% of the oracle ceiling on the 34 tasks where it is identified; three nonlinear readouts do no better, indicating that readout capacity is not the bottleneck. Headline results use one ensemble per task, with variation measured across episode splits. The growth-rate, rather than asymptotic, form of Theorem 4 applies here, and because error is flat on 21 of 25 Atari games, evidence for gains under compounding error comes from DMC and TD-MPC2. The intervals require exchangeable calibration and test rollouts and support trusted-horizon decisions, but neither uncertainty penalties nor truncation improved real return, while a preregistered safety filter passed on only 2 of 13 tasks. Using the intervals for planning would therefore require calibration on the planner's own rollout distribution, as indicated by our behavior-shift results (Appendix H).

7 Conclusion

We prove that recurrent state-space world-model latents are compact, imposing a horizon-independent ceiling on model-internal uncertainty. Empirically, ensemble disagreement saturates on most DreamerV3 tasks with compounding reward error and becomes anti-correlated with error on 7 of 19 TD-MPC2 tasks. We address this without retraining or ensembles. A linear probe of the frozen latent, calibrated by split conformal prediction at each horizon, achieves the 90% coverage target across both model families and improves sharpness where error compounds, outperforming stabilized ensemble scales on 14 of 17 DreamerV3 tasks. The calibrated intervals provide a reliable trusted-horizon signal, accepting as many imagined steps as ensemble disagreement while violating the error tolerance less often. However, they do not improve closed-loop planning, narrow joint multi-horizon bands, or remain calibrated under behavior-policy shifts without recalibration. Thus, the latent retains useful ordering information despite its bounded magnitude, while conformal calibration supplies the missing scale. Why disagreement saturates remains open, as both mechanisms we tested were falsified. A hypothesis we did not test is the training objective, which differs sharply across the two families, reconstruction with a one-step KL term in DreamerV3 (Hafner et al., 2024) versus a multi-step, decision-aligned target in TD-MPC2 (Hansen et al., 2024). Varying only this objective at fixed architecture would test whether the ceiling and the saturation pattern move with it.

AI USE STATEMENT

We used AI assistants (large language models) during the preparation of this manuscript for drafting and editing prose, formatting LaTeX, and reviewing code for correctness. All experimental results were generated by code we wrote and executed ourselves; no AI system produced, modified, or verified any experimental data or numerical claim. All theoretical proofs were conceived and verified by the authors. The AI tools assisted manuscript preparation only; the authors take full responsibility for the scientific content of this work.

REPRODUCIBILITY STATEMENT

All code is included in the supplementary material, and the pipeline is deterministic given a set of seeds. Full details are in Appendix J.

REFERENCES

Donald Aadithiyan. Reading the room: Implicit confusion encoding in recurrent world model states, 2026. URL https://arxiv.org/abs/2608.21582.

Mehran Aghabozorgi, Alireza Moazeni, Yanshu Zhang, and Ke Li. Wimle: Uncertainty-aware world models with imle for sample-efficient continuous control, 2026. URL https://arxiv.org/abs/2602.14351.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2021. URL https://arxiv.org/abs/2107.07511.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture, 2023. URL https://arxiv.org/abs/2301.08243.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation, 2013.

Julia Berger, Bernd Frauenknecht, Sebastian Trimpe, and Bastian Leibe. Biased dreams: Limitations to epistemic uncertainty quantification in latent dynamics models, 2026. URL https://arxiv.org/abs/2604.25416.

Lukas Bierling, Davide Pasero, Jan-Henrik Bertrand, and Kiki Van Gerwen. Dreamerv3-xp: Optimizing exploration through uncertainty estimation, 2025. URL https://arxiv.org/abs/2510.21418.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models, 2018. URL https://arxiv.org/abs/1805.12114.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Christopher Grimm, André Barreto, Satinder Singh, and David Silver. The value equivalence principle for model-based reinforcement learning, 2020. URL https://arxiv.org/abs/2011.03506.

Danijar Hafner. Benchmarking the spectrum of agent capabilities, 2022. URL https://arxiv.org/abs/2109.06780.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models, 2021. URL https://arxiv.org/abs/2010.02193.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models, 2024. URL https://arxiv.org/abs/2301.04104.

Nicklas Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control, 2024. URL https://arxiv.org/abs/2310.16828.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization, 2019. URL https://arxiv.org/abs/1906.08253.

Lukasz Kaiser, Mohammad Babaeizadeh, Piotr Milos, Blazej Osinski, Roy H Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, Afroz Mohiuddin, Ryan Sepassi, George Tucker, and Henryk Michalewski. Model-based reinforcement learning for atari, 2020. URL https://arxiv.org/abs/1903.00374.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles, 2017. URL https://arxiv.org/abs/1612.01474.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B*, 76(1):71–96, 2014.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression, 2018. URL https://arxiv.org/abs/1604.04173.

Chenhao Li, Andreas Krause, and Marco Hutter. Uncertainty-aware robotic world model makes offline model-based reinforcement learning work on real robots, 2026. URL https://arxiv.org/abs/2504.16680.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression, 2019. URL https://arxiv.org/abs/1905.03222.

Annika Schneider, Tommy Rochussen, Joshua Stiller, and Vincent Fortuin. Decision-aligned evaluation of uncertainty quantification, 2026. URL https://arxiv.org/abs/2606.26990.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model, 2020. URL https://arxiv.org/abs/1911.08265.

Tim Seyde, Wilko Schwarting, Sertac Karaman, and Daniela Rus. Learning to plan optimistically: Uncertainty-guided deep exploration via latent model ensembles, 2021. URL https://arxiv.org/abs/2010.14641.

Kamilė Stankevičiūtė, Ahmed M. Alaa, and Mihaela van der Schaar. Conformal time-series forecasting. In *Advances in Neural Information Processing Systems*, 2021.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. Deepmind control suite, 2018. URL https://arxiv.org/abs/1801.00690.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, 2019.

Claas Voelcker, Victor Liao, Animesh Garg, and Amir massoud Farahmand. Value gradient weighted model-based reinforcement learning, 2023. URL https://arxiv.org/abs/2204.01464.

Claas Voelcker, Anastasiia Pedan, Arash Ahmadian, Romina Abachi, Igor Gilitschenski, and Amir massoud Farahmand. Calibrated value-aware model learning with probabilistic environment models, 2025. URL https://arxiv.org/abs/2505.22772.

Claas A Voelcker, Arash Ahmadian, Romina Abachi, Igor Gilitschenski, and Amir massoud Farahmand. λ-models: Effective decision-aware reinforcement learning with latent models, 2024. URL https://arxiv.org/abs/2306.17366.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Stefan Radic Webster and Peter Flach. Risk sensitive model-based reinforcement learning using uncertainty guided planning, 2021. URL https://arxiv.org/abs/2111.04972.

Hongzhan Yu, Chenghao Li, Ruipeng Zhang, Henrik Christensen, and Sicun Gao. Sensitivity shaping for latent modeling, 2026a. URL https://arxiv.org/abs/2606.14585.

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

A PROOFS

Proof of Proposition 1. By induction. The initial state is $h_0 = \mathbf{0} \in [-1, 1]^D$. Assume $h_{t-1} \in [-1, 1]^D$. Componentwise, $u_{t,i} = \sigma(\cdot) \in (0, 1)$ and $c_{t,i} = \tanh(\cdot) \in (-1, 1)$, so $h_{t,i} = u_{t,i}\, c_{t,i} + (1 - u_{t,i})\, h_{t-1,i}$ is a convex combination of two values in $[-1, 1]$, hence in $[-1, 1]$. □

Proof of Proposition 2. Each of the S groups is a one-hot vector, via the straight-through estimator (Bengio et al., 2013), which is numerically one-hot in the forward pass, so $\|z_t\|_2^2 = S$. Two one-hot vectors over C classes differ in at most one position, which gives a per-group distance of $\sqrt{2}$. Summing over S groups yields $\sqrt{2S}$. □

Proof of Corollary 3. $\|h_t - h'_t\|_2 \le 2\sqrt{D}$, since both lie in $[-1, 1]^D$, and $\|z_t - z'_t\|_2 \le \sqrt{2S}$ by Proposition 2. The joint bound follows from $\sqrt{(2\sqrt{D})^2 + (\sqrt{2S})^2} = \sqrt{4D + 2S}$. □

Proof of Proposition 5. Condition on the fitted scale and on the train episodes that produced it. Conditional on them, the n calibration ratios and the test ratio remain exchangeable, since the calibration and test episodes were never used in the fit. With exchangeable random variables the test ratio is equally likely to fall at any of the $n + 1$ ranks among the pooled scores. The interval contains the test point exactly when its rank among the pooled scores is at most $k = \lceil (n + 1)(1 - \alpha) \rceil$, and $P(\text{rank} \le k) = k/(n + 1) \ge 1 - \alpha$. □

Proof of Theorem 4. The bound $U_t \le U_{\max}$ follows from Corollary 3 and the Lipschitz assumption. If $\sup_t \delta_t = \infty$, then for any finite $U_{\max}$ there is a t^* with $\delta_{t^*} > U_{\max}/L$. For $t \ge t^*$ the range $[0, U_{\max}]$ cannot distinguish trajectories whose physical errors differ by more than $U_{\max}/L$. □

Corollary 6 (Decoding does not escape the bound)**.** *Let $K = [-1, 1]^D \times \mathcal{Z}$ be the latent state set, which is compact by Propositions 1 and 2. Let $g : K \to \mathbb{R}^m$ be any continuous function of the latent, such as the decoder, the reward head, the value head or any ensemble of these. Then $g(K)$ is a continuous image of a compact set, so it is compact and bounded by a finite M_g that depends only on the trained weights. Any uncertainty statistic that is a Lipschitz function of the model's own outputs is therefore bounded, whatever space it is read out in.*

Proof. A continuous image of a compact set is compact, hence bounded, by the extreme value theorem. □

B FULL EXPERIMENTAL RESULTS

B.1 EVERY ROW

Table 3 lists every task in the study with its coverage, the late-horizon adaptivity of the probe and of each conformal baseline, and the probe's interval score relative to plain conformal. It is generated from the result files by `analysis/paper_tables.py`, as is every other table in this paper.

Table 3: All 60 rows: coverage, late-horizon adaptivity for the probe, ensemble-scaled, k-NN and Mondrian conformal, and the probe's interval score relative to plain conformal.

Family	Task	Cov.	Probe	Ens	k-NN	Mond	IS/conf
DV3	acrobot_swingup	0.908	+0.842	+0.565	+0.816	+0.580	0.49
DV3	cartpole_swingup	0.907	+0.990	+0.889	+0.979	+0.444	0.76
DV3	cheetah_run	0.905	+0.704	+0.067	+0.600	+0.053	0.96
DV3	cup_catch	0.908	+0.953	+0.677	+0.941	+0.093	0.97
DV3	finger_spin	0.907	+0.891	−0.789	+0.777	+0.554	0.93
DV3	finger_turn_hard	0.909	+0.547	+0.002	+0.315	−0.152	1.05
DV3	hopper_hop	0.906	+0.866	+0.236	+0.838	+0.469	0.61
DV3	reacher_easy	0.907	+0.573	+0.867	+0.623	+0.476	1.01
DV3	reacher_hard	0.905	+0.621	−0.215	+0.342	+0.046	1.00
DV3	walker_run	0.909	+0.876	−0.028	+0.851	+0.043	0.94
DV3	walker_walk	0.907	+0.863	−0.540	+0.590	+0.513	0.92
DV3	cartpole_swingup	0.908	+0.986	+0.133	+0.961	−0.175	0.68
DV3	cheetah_run	0.906	+0.930	−0.183	+0.752	+0.102	0.87
DV3	hopper_hop	0.905	+0.972	+0.087	+0.943	+0.238	0.44
DV3	walker_run	0.908	+0.950	−0.516	+0.618	+0.464	0.92
DV3	alien	0.909	+0.140	+0.266	+0.109	−0.006	0.96
DV3	amidar	0.906	+0.141	+0.110	+0.082	+0.019	1.00
DV3	assault	0.907	+0.375	+0.320	+0.133	+0.187	0.97
DV3	asterix	0.912	+0.340	+0.285	+0.227	+0.146	0.90
DV3	bank_heist	0.904	+0.367	+0.189	+0.191	+0.063	1.00
DV3	battle_zone	0.901	+0.140	+0.164	+0.082	+0.033	1.00
DV3	boxing	0.906	+0.433	+0.308	+0.217	+0.150	0.73
DV3	breakout	0.914	+0.246	+0.270	+0.037	+0.081	0.99
DV3	chopper_command	0.906	+0.181	+0.253	+0.081	+0.117	0.99
DV3	crazy_climber	0.902	+0.795	+0.513	+0.564	+0.402	0.50
DV3	demon_attack	0.906	+0.066	+0.097	+0.008	+0.030	1.00
DV3	frostbite	0.909	+0.542	+0.467	+0.180	+0.180	0.85
DV3	gopher	0.903	+0.347	+0.205	+0.116	+0.194	0.97
DV3	hero	0.907	+0.733	+0.758	+0.703	+0.402	0.58
DV3	jamesbond	0.909	+0.341	+0.478	+0.094	+0.105	1.00
DV3	kangaroo	0.907	+0.419	+0.438	+0.144	+0.107	0.99
DV3	krull	0.903	+0.611	+0.292	+0.528	+0.256	0.76
DV3	kung_fu_master	0.902	+0.537	+0.249	+0.311	+0.214	1.00
DV3	ms_pacman	0.908	+0.133	+0.293	+0.105	−0.020	0.97
DV3	pong	0.906	+0.618	+0.674	+0.313	+0.272	0.80
DV3	private_eye	0.902	+0.131	+0.612	+0.212	+0.082	1.00
DV3	qbert	0.909	+0.678	+0.083	+0.578	+0.017	0.57
DV3	road_runner	0.906	+0.626	+0.099	+0.322	+0.054	0.98
DV3	seaquest	0.904	+0.234	+0.250	+0.094	+0.115	1.00
DV3	up_n_down	0.901	+0.332	+0.289	+0.163	+0.096	0.97
TD	acrobot_swingup	0.905	+0.780	+0.842	+0.683	+0.426	0.62
TD	cheetah_run	0.903	+0.624	+0.284	+0.437	+0.185	0.93
TD	cup_catch	0.906	+0.959	+0.930	+0.913	+0.347	0.98
TD	finger_spin	0.903	+0.608	−0.977	+0.397	+0.463	1.10
TD	finger_turn_hard	0.901	+0.695	+0.659	+0.348	+0.474	0.78
TD	reacher_easy	0.900	+0.330	+0.850	+0.223	+0.573	0.92
TD	reacher_hard	0.893	+0.632	+0.804	+0.413	+0.415	0.75
TD	walker_run	0.904	+0.777	−0.778	+0.658	+0.622	1.00
TD	walker_walk	0.904	+0.881	−0.662	+0.753	+0.605	0.96
TD	acrobot_swingup	0.902	+0.832	+0.880	+0.714	+0.350	0.74
TD	cartpole_swingup	0.907	+0.977	+0.628	+0.952	+0.547	0.31
TD	cheetah_run	0.902	+0.757	+0.580	+0.610	+0.536	0.68
TD	cup_catch	0.905	+0.937	+0.674	+0.945	+0.153	0.99
TD	finger_spin	0.904	+0.692	−0.277	+0.417	+0.033	1.21
TD	finger_turn_hard	0.893	+0.240	+0.014	+0.175	−0.018	1.34
TD	reacher_easy	0.897	+0.780	+0.265	+0.624	+0.130	0.74
TD	reacher_hard	0.900	+0.760	−0.043	+0.504	−0.021	0.65
TD	walker_run	0.905	+0.916	+0.678	+0.796	+0.602	0.87
TD	walker_walk	0.906	+0.982	−0.375	+0.924	+0.130	0.90
Cr	crafter_reward	0.912	+0.431	+0.256	+0.150	+0.144	0.77

C Does the Ceiling Bind Per Rollout?

Theorem 4 bounds a latent statistic, so one might expect the spread of disagreement across rollouts to stop growing while the spread of error keeps growing. We measured that directly, comparing the $p_{95} - p_5$ spread of each quantity early ($h < 10$) against late ($h \geq H/2$). On the 36 tasks where error compounds, the median error spread grows $2.00\times$ and the median disagreement spread $1.38\times$, and disagreement's spread grows more slowly on 21 of the 36. The extremes are stark, with multi-task reacher_hard showing an error spread growing $1512\times$ against $1.17\times$ for disagreement, but a majority of 21 of 36 is weak. We therefore report the ceiling as binding on the growth of the means, which is what the saturation classification measures, and not as a clean statement about per-rollout spread. This measurement is descriptive and was made after the fact.

D Discussion of the Bound

What the Lipschitz class covers. Ensemble disagreement over predicted latents is the spread of K embeddings of the same state, bounded by the same divergence through the triangle inequality; the prior KL and the predictive variance of any bounded head are functions of the same compact quantity; and Corollary 6 extends the argument to statistics read after decoding. The latent of TD-MPC2, a product of simplices under SimNorm, is compact for the same reason (Appendix F.3), so the ceiling is a property of the design pattern, not of one architecture. The proposition is the other half of the story: once a scale is chosen, the conformal guarantee is indifferent to it, so all the design weight falls on ordering, and ordering is what the ceiling leaves intact.

Practicality of the bound. Theorem 4 is asymptotic, since it requires $\sup_t \delta_t = \infty$, which is doubtful on DMC tasks with bounded state spaces. The version that applies to our benchmarks is about growth rates. Latent divergence reaches its effective ceiling almost immediately: we measure it at 6.65, flat to three digits from $h = 1$ to $h = 50$, which is 14.5% of the worst case 45.96. Over the horizons where the latent statistic is already pinned, physical error keeps growing: a constant statistic carries no information about a changing error. This is a measured regularity rather than an established mechanism. We tested two explanations and falsified both. The deterministic update is not contractive along trained trajectories, since its Jacobian spectral radius is 1.00 to 1.14 on 9 of 11 DMC proprioceptive tasks, including finger_spin, where the inversion is strongest. We also pre-registered the prediction that imagined rollouts lose reward diversity faster on saturating tasks than on tasks where disagreement keeps pace, and tested it over 29 tasks; it failed (permutation $p = 0.34$, Appendix F.2).

E Experimental Details

E.1 World Model Configuration

We use the official DreamerV3 implementation pinned at commit `e3f0224`. DMC proprioceptive tasks use the `size1m` configuration, with deterministic dimension $D = 512$, stochastic groups $S = 32$, classes per group $C = 4$, uniform mixture $\eta = 0.01$ and RMSNorm; DMC pixel tasks, Atari and Crafter use the default size, whose latent is 10240-dimensional. DMC tasks are trained for 1M environment steps and Atari 100k tasks for 100k. Ensembles use $K = 5$ independent training runs with different seeds on most DMC proprioceptive tasks and $K = 3$ elsewhere. TD-MPC2 uses the released single-task and multi-task checkpoints, with 3 and 5 seeds respectively. We collect 40 evaluation episodes per DreamerV3 task, 94 for Crafter, and 80 (single-task) or 60 (multi-task) for TD-MPC2.

E.2 Probe Hyperparameters

The ridge penalty λ is selected per task on calibration data by `method/tune_probe.py`, searching over $\lambda \in \{0.001, 0.01, 0.1, 1.0, 10.0, 100.0\}$. The scale stabilizer $\beta \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1.0\}$ is selected on train episodes to minimize mean interval width. The k-NN neighborhood $k \in \{5, 10, 20, 40\}$ is selected by the same rule. The conformal target is $\alpha = 0.1$ (90% coverage). Every estimator behind the headline table, the budget table and the per-task rows uses 120 random three-way episode splits. Three auxiliary analyzes use 60: the joint multi-horizon bands, the label-efficiency sweep, and the runs at $\alpha = 0.05$ and $\alpha = 0.2$.

E.3 Hardware and Software

Training and evaluation were conducted on the NVIDIA A100 GPUs. Software: Python 3.10, JAX 0.4.x, NumPy 1.24, the official DreamerV3 codebase.

F Further Comparisons

Replication across retrained models. We retrained walker_walk, finger_spin and reacher_easy and formed three disjoint 3-seed ensembles per task, so each group is an independently trained world model. Adaptivity moves between groups, by a spread up to $4.3\times$ larger than the split-resampling standard deviations suggest, and reacher_easy is the extreme case, where the probe scores $+0.568$, $+0.454$ and $+0.083$ across groups. The per-task verdict is stable. The probe beats ensemble, k-NN and Mondrian scaling in every group on walker_walk and finger_spin, and in none on reacher_easy,

Table 4: Every task where ensemble disagreement ends up anti-correlated with the error it should track. `Crossing` is the first horizon at which the correlation reaches zero. Saturating marks a task where disagreement grows more slowly than error.

Family	Task	Early	Late	Crossing	Saturating
TD-MPC2 proprio	finger_spin	$+0.751$	-0.982	10	yes
TD-MPC2 proprio	walker_walk	$+0.883$	-0.895	27	yes
DV3 proprio	finger_spin	-0.383	-0.893	2	yes
TD-MPC2 proprio	walker_run	$+0.886$	-0.886	24	yes
DV3 proprio	walker_walk	$+0.891$	-0.791	28	yes
DV3 vision	walker_run	$+0.214$	-0.521	1	yes
7 further tasks invert; the full list is in Table 3					

Table 5: The probe against conformalized quantile regression on adaptivity, with CQR's own coverage and the width each pays relative to plain conformal. The two-sided exact binomial sign test is used here rather than the one-sided value used elsewhere.

	DreamerV3	TD-MPC2
Probe more adaptive	$38/39$	$19/19$
Compounding rows	$17/17$	$19/19$
CQR coverage ok	$39/39$	$19/19$
p (two-sided exact sign)	1.5×10^{-10}	3.8×10^{-6}
Median adaptivity (probe / CQR)	$+0.573$ / $+0.179$	$+0.777$ / $+0.307$
Median width (probe / CQR)	$0.85\times$ / $2.33\times$	$0.81\times$ / $1.47\times$

matching what the full ensembles say. On interval score the `dis2` probe beats the stabilized ensemble in 8 of the 9 groups, against a pre-registered bar of 7, and the exception is a reacher_easy group where the ensemble reaches 0.353 times plain conformal.

Quantile regression as the supervised baseline. Table 5 reports the comparison against conformalized quantile regression, the one baseline that also fits a supervised model of the residual. CQR holds coverage on every row it ran, so this is a valid, correctly implemented baseline losing on ordering, not a broken one: pooled over both families the probe is more adaptive on 57 of 58 settings (two-sided exact sign test, $p = 4 \times 10^{-16}$), and CQR pays $2.33\times$ the plain-conformal width on DreamerV3 for its adaptivity where the probe pays $0.85\times$. One row is absent from the count: on atari100k_up_n_down the dump carries 3816 rollouts against roughly 533 for its peers and CQR did not converge there in 61 CPU-days.

F.1 Supervision Is the Discriminator

Table 6 compares the supervised probe against four unsupervised signals. These are deep ensemble disagreement, cumulative ensemble disagreement, latent norm, and the single model's predictive standard deviation of reward. The supervised probe is the best signal on 21 of the 40 DreamerV3 tasks, and on 33 of the 60 tasks pooled over all three families. An unsupervised signal beats it on 19 tasks, but 18 of those are Atari games where the probe scores near zero, which is again the regime where error does not grow and no signal is informative. On the 15 DMC proprioceptive and vision tasks the probe is the best signal on 14. The single loss is reacher_easy, where deep ensemble disagreement reaches $+0.922$ against the probe's $+0.789$.

The pattern is consistent. Supervision wins where there is error to track and says nothing where there is not. The unsupervised signals that occasionally beat the probe on Atari do so at correlations between $+0.1$ and $+0.3$, too weak to produce a useful adaptive interval for any method. This matches the theoretical reading, in which what separates signals that work from signals that fail is supervision rather than boundedness.

F.2 The Saturation Mechanism

Table 7 and Figure 6 examine growth-rate saturation. For each task we compute a growth ratio, the factor by which disagreement grows from early to late horizons divided by the factor by which error grows. A ratio below 1 means disagreement grows more slowly than error, which is what we

Table 6: Late-horizon correlation with true error, for the supervised probe against unsupervised signals. The supervised probe is best on 21 of 40 DreamerV3 tasks and 14 of 15 on DMC; pooled over all three families the table below reports 33 of 60. Losses are concentrated on Atari where error is flat.

Family	Task	Deep ens	Cum ens	Latent norm	Rew std	Probe
DV3 proprio	acrobot_swingup	$+0.566$	$+0.498$	$+0.201$	$+0.308$	$+0.824$
DV3 proprio	cartpole_swingup	$+0.886$	$+0.856$	-0.775	-0.022	$+0.991$
DV3 proprio	cheetah_run	$+0.112$	$+0.180$	-0.187	-0.067	$+0.830$
DV3 proprio	cup_catch	$+0.692$	$+0.501$	$+0.834$	$+0.171$	$+0.978$
DV3 proprio	finger_spin	-0.799	-0.431	$+0.111$	-0.086	$+0.890$
DV3 proprio	finger_turn_hard	$+0.218$	$+0.217$	-0.251	-0.080	$+0.453$

The probe is the best signal on 33 of 60 tasks with all five signals scored.

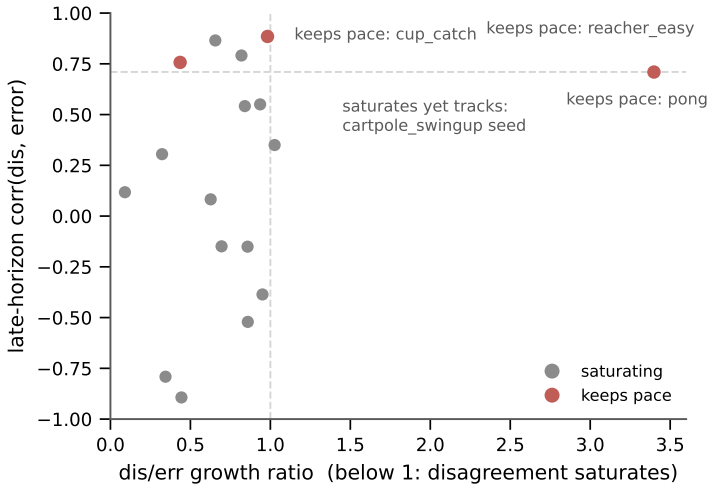

Figure 6: Growth-rate saturation against late-horizon tracking. Each point is one of the 17 answerable DreamerV3 rows, one per ensemble, the rows of Table 7. The x-axis is the ratio of disagreement growth to error growth, and the y-axis is the late-horizon correlation between disagreement and true error. Every task where disagreement keeps pace tracks the error, and most saturating rows do not, with exceptions. The association is descriptive: a pre-registered mechanism test failed (Section F.2).

call saturation. On the 17 DreamerV3 tasks where error grows significantly under an episode-level bootstrap (the "answerable" tasks), disagreement saturates on 14.

The relationship is one-directional. The three tasks where disagreement keeps pace (reacher_easy, cup_catch, pong) all retain late correlation above $+0.70$, and these are the tasks where ensemble-scaled conformal remains competitive. Saturation usually, but not always, breaks tracking: 12 of the 14 saturating tasks fall below $+0.71$, but cartpole_swingup saturates and still tracks at $+0.865$. The 23 flat-error tasks cannot speak to the question.

The regularity does not transfer to TD-MPC2. There, two of the ten tasks where disagreement keeps pace have late correlation near zero, and inversions occur in both classes. Nor does it come with a mechanism. We pre-registered a test of whether imagined rollouts lose reward diversity faster on saturating tasks, over 29 tasks from both families. All three predictions failed. Collapse gave a permutation $p = 0.34$, its correlation with late tracking gave Spearman -0.09, and the timing of latent drift gave $p = 0.25$. A contraction account is ruled out separately (Section 4). We therefore report saturation as a measured description of when disagreement fails on DreamerV3, not as an explanation of why.

Table 7: Growth-rate saturation, showing whether disagreement keeps pace with error. `dis/err` is disagreement growth divided by error growth; the class column comes from a growth-difference bootstrap on the same rollouts, so a row can keep pace by the test while its point ratio sits below 1. On DreamerV3, every task where disagreement keeps pace tracks the error (3 of 3), and 12 of the 14 saturating rows fall below $+0.71$.

Family	Task	Err growth	Dis growth	Dis/Err	Late corr	Class
TD	finger_spin	$3.63\times$	$0.13\times$	0.03	-0.982	saturating
TD	finger_spin	$16.80\times$	$0.92\times$	0.05	-0.457	saturating
DV3	hopper_hop	$0.87\times$	$0.08\times$	0.09	$+0.118$	saturating
TD	walker_walk	$71.40\times$	$10.12\times$	0.14	-0.414	saturating
TD	walker_run	$155.46\times$	$28.07\times$	0.18	-0.886	saturating
TD	reacher_easy	$50.10\times$	$11.10\times$	0.22	$+0.267$	keeps pace
TD	walker_walk	$205.82\times$	$48.28\times$	0.23	-0.895	saturating
TD	walker_run	$46.26\times$	$13.88\times$	0.30	$+0.654$	saturating
DV3	hopper_hop	$1.51\times$	$0.49\times$	0.32	$+0.305$	saturating
DV3	walker_walk	$7.39\times$	$2.54\times$	0.34	-0.791	saturating
TD	acrobot_swingup	$157.28\times$	$64.49\times$	0.41	$+0.867$	saturating
DV3	cup_catch	$2.03\times$	$0.88\times$	0.44	$+0.757$	keeps pace
DV3	finger_spin	$1.44\times$	$0.64\times$	0.44	-0.893	saturating
TD	cheetah_run	$10.33\times$	$5.62\times$	0.54	$+0.177$	saturating
DV3	cartpole_swingup	$1.80\times$	$1.13\times$	0.63	$+0.082$	saturating
DV3	cartpole_swingup	$1.69\times$	$1.11\times$	0.66	$+0.865$	saturating
TD	acrobot_swingup	$36.94\times$	$25.62\times$	0.69	$+0.878$	keeps pace
DV3	walker_run	$1.49\times$	$1.03\times$	0.69	-0.149	saturating
TD	reacher_hard	$29.63\times$	$21.03\times$	0.71	-0.044	keeps pace
DV3	hero	$0.99\times$	$0.81\times$	0.82	$+0.791$	saturating
TD	cheetah_run	$6.72\times$	$5.54\times$	0.82	$+0.608$	saturating
DV3	acrobot_swingup	$3.32\times$	$2.79\times$	0.84	$+0.542$	saturating
TD	finger_turn_hard	$9.45\times$	$7.95\times$	0.84	$+0.655$	keeps pace
DV3	finger_turn_hard	$1.23\times$	$1.06\times$	0.86	-0.151	saturating
DV3	walker_run	$1.24\times$	$1.07\times$	0.86	-0.521	saturating
DV3	krull	$1.10\times$	$1.03\times$	0.94	$+0.550$	saturating
DV3	reacher_hard	$1.09\times$	$1.04\times$	0.95	-0.386	saturating
DV3	reacher_easy	$3.39\times$	$3.33\times$	0.98	$+0.885$	keeps pace
TD	cartpole_swingup	$12.32\times$	$12.38\times$	1.01	$+0.516$	keeps pace
DV3	crazy_climber	$1.22\times$	$1.26\times$	1.03	$+0.350$	saturating
TD	reacher_hard	$15.74\times$	$16.79\times$	1.07	$+0.831$	keeps pace
TD	reacher_easy	$7.72\times$	$8.67\times$	1.12	$+0.897$	keeps pace
TD	cup_catch	$4.28\times$	$5.78\times$	1.35	$+0.941$	keeps pace
TD	cup_catch	$2.52\times$	$4.00\times$	1.59	$+0.698$	keeps pace
DV3	pong	$2.92\times$	$9.93\times$	3.40	$+0.710$	keeps pace
TD	finger_turn_hard	$1.52\times$	$6.28\times$	4.12	-0.032	keeps pace

F.3 Is Normalization the Mechanism?

The method assumes only a frozen latent state and measured errors, and transfers unchanged to TD-MPC2, whose SimNorm latent is a product of simplices and therefore also compact. What does not transfer is a general account of why latent uncertainty fails. It is tempting to blame the normalization that makes these models trainable, whether categorical codes, SimNorm or sphere normalization. We tested that with two small JEPA-style models identical except for L2 normalization, and neither reproduced the failure, although compactness did cut the level of the latent's own disagreement-error correlation about threefold. Those models were small, so the account is not ruled out at scale, but we do not make it. The bound says why a latent statistic cannot report the magnitude of the error, not which models or tasks will invert.

F.4 Where the Coverage Goes

Marginal coverage is a tie between any two valid split-conformal methods, so it cannot separate them; what separates them is how that coverage is distributed across rollout difficulty. Plain conformal gives every rollout the same width, so its coverage varies with difficulty. The probe varies the width instead and holds coverage closer to the target in both tails (Figure 7): on walker_walk it

covers the easy and hard thirds at 0.929 and 0.875, against 0.995 and 0.762 for plain conformal, at matched marginal coverage (0.909 and 0.910). Stratifying by the realized error conditions on the response, so the level in either tail is not interpretable against the 90% target; what the comparison isolates is the uniformity of coverage across difficulty at matched marginal coverage.

That uniformity does not depend on how difficulty is defined. Across the 11 DMC proprioceptive tasks we repeated the comparison under two stratifiers that do not use the outcome: the rollout's error over $h \leq 5$, which a deployed agent observes directly, and ensemble disagreement. The probe's gap between easy- and hard-third coverage is unchanged at 0.054, 0.057 and 0.056 under the three definitions, while plain conformal's falls from 0.214 to 0.105 to 0.074, tracking how much the stratifier knows about the error. A constant-width interval is uniform only along axes that carry no information about difficulty. Under disagreement stratification the stabilized ensemble has the largest gap of the three arms, 0.116, so that scale is not uniform even along the axis it is built from.

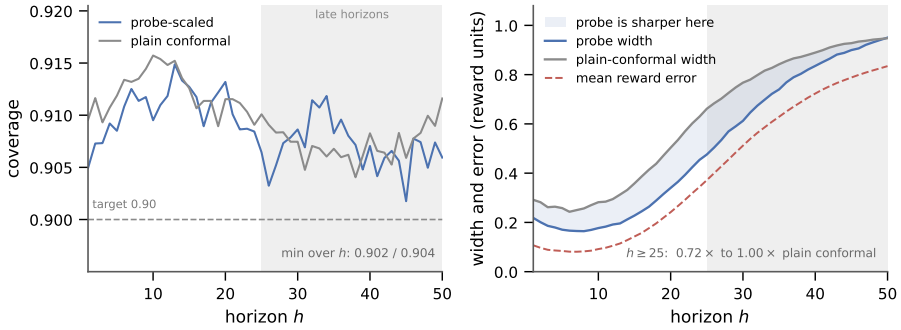

Figure 7: Coverage (left) and interval width (right) by horizon on walker_walk, with rollouts split into the easiest and hardest third by mean reward error; each band spans the two. Plain conformal's width band collapses to a single line and its coverage band does not.

F.5 Probe Variants and Ablation

We compare the latent-only probe against the richer `latent+h+dis2` variant, which adds the normalized horizon index, ensemble disagreement and its square to the probe input. Giving the probe the unsupervised signal helps where that signal is informative. On adaptivity, the richer variant beats the stabilized ensemble scale on 35 of the 40 DreamerV3 tasks against 27 for the latent-only probe, and on 18 of 19 TD-MPC2 tasks against 15. On interval score it wins 23 of 40 against 22, and 15 of 17 of the compounding tasks against 14, with 13 of 19 on TD-MPC2 against 10.

Nonlinear readouts of the same latent do not help, and the comparison has to be made against a matched control. On the 40 tasks whose latent is 10240-dimensional the richer variants are fitted on a 256-dimensional PCA, so the linear probe fitted on that same reduction, not the full-latent probe, is what isolates the nonlinearity. Random Fourier features lose against it on 39 of those 40 tasks and a per-horizon fit on 37, at median adaptivity changes of -0.110 and -0.085. A quadratic expansion is close to a wash, winning 10 of the 40 at a median change of -0.007. On the 20 tasks that need no reduction, where the two controls coincide, the picture is the same: 1 of 20 for random Fourier features, 5 of 20 per-horizon, and 9 of 20 for the quadratic expansion at a median change of -0.001. No variant improves the median on any subset, so readout capacity is not what limits the probe, but a quadratic expansion of the same features is a coin flip rather than a clear loss.

The tasks where raw disagreement still wins are flat-error Atari games, where both signals score near zero. The clearest case in the other direction is reacher_easy, the one task where disagreement clearly beats the latent-only probe. There the probe scores $+0.569$ and rises to $+0.900$ once disagreement is an input, so the probe absorbs the unsupervised signal rather than replacing it. The same variant can exceed the oracle ceiling computed on the latent alone, which is expected, since disagreement is not linearly present in the latent.

F.6 Published Uncertainty Signals

Two further families of single-model uncertainty need no ensemble. We evaluated the latent-transition disagreement of Berger et al. (2026) and a last-layer Laplace approximation of the reward head on the 11 DMC proprioceptive tasks, in both cases as the conformal scale for a single

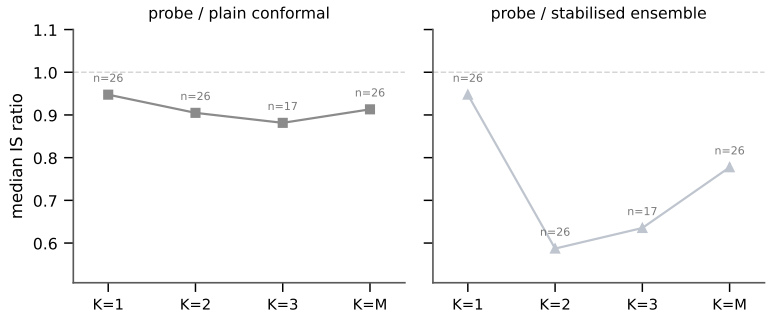

Figure 8: Interval score of the probe relative to plain conformal (left) and to the stabilized ensemble scale (right) as the ensemble size varies over $K \in \{1, 2, 3, M\}$, with the number of scored rows annotated. At $K{=}1$ the ensemble scale degenerates to plain conformal, so the right panel starts from the same point as the left. The pre-registered bars of Table 8 pass at every size tested. The right panel is non-monotone, and the cause is the comparator rather than the probe. The probe's own interval score improves monotonically with K, from 0.927 at $K = 1$ to 0.795 at $K = M$ in median over these settings, while the ensemble scale is worst at $K = 2$, where two members give the noisiest disagreement estimate, and only partly recovers by $K = M$. It never returns to the level of plain conformal, so the gap the probe opens over it does not close.

Table 8: Pre-registered bars for the ensemble-size question, written down before the run that tested them. K1 is required to pass at every K rather than on average; K3 has seven rows, so it is a consistency check, not a power claim.

Bar	Result	Verdict
K1: sharper than plain at every K	at $K{=}1/2/3$: $20/26$, $20/26$, $14/17$	PASS
K2: median ratio at $K{=}2 \leq$ at $K{=}M$	0.587 vs 0.778	PASS
K3: verdict agrees with $K{=}M$ ($M \geq 5$)	$7/7$	PASS

checkpoint. The Laplace baseline uses both the Monte Carlo and the linearized predictive, with prior precision set by marginal likelihood. The probe has the lower interval score and the higher late adaptivity than every variant of both baselines on all 11 tasks. The Laplace scale is worse than a constant one on the median task, with an interval score between 1.15 and 3.95 times plain conformal and mean late adaptivity no higher than $+0.17$. The model's own predictive spread of reward has mean late adaptivity -0.04. Against MOPO-style disagreement the probe has the lower interval score on 28 of the 40 DreamerV3 tasks, including all 17 where error compounds, and on all 9 TD-MPC2 tasks where it is defined. MOReL-style disagreement behaves like the ensemble scale discussed above.

Other confidence levels. Re-scoring all 62 evaluation settings at $\alpha = 0.05$ and again at $\alpha = 0.2$ leaves the guarantee intact: the probe lands within ± 0.02 of its target on 62 of 62 rows at both levels, and 48 of the 62 rows hold target coverage and beat plain conformal at both. The two rows that miss at every level are the TD-MPC2 finger_spin rows, already the weakest at $\alpha = 0.1$, so the exceptions are consistent rather than random.

F.7 Ensemble Size, One Checkpoint, and Width Stability

Ensemble size. Table 8 reports the three pre-registered bars, and Figure 8 the interval-score ratios behind them, re-scored on K-member subsets rebuilt from the per-member dumps over 31 scored rows (19 DreamerV3, 10 TD-MPC2 and the two nine-seed re-dumps). The $K{=}M$ references come from the same dump as the subsets; the two dump collections behind a row differ by up to 0.85 in measured error, so a reference read from the other collection would compare the two arms on different rollouts.

One checkpoint. Table 9 reports the probe fitted on a single checkpoint's own latent against that checkpoint's own error, on the 29 rows with per-member dumps, with the regularization re-selected so nothing is inherited from the ensemble-mean fit. Validity does not depend on the ensemble at all:

Table 9: The probe on one checkpoint ($K=1$), with no ensemble anywhere. Bar rows carry the pre-registered threshold and its verdict; the remaining rows are descriptive context.

Quantity	Value	Bar
Rows within ± 0.02 of 0.90 coverage	$29/29$	$\geq 95\%$
Coverage range	0.8945 to 0.9098	–
Hard-third coverage beats plain conformal	$22/29$	$\geq 2/3$
Mean hard-third coverage	probe 0.830 vs plain 0.747	–
Median interval score / plain, compounding rows	0.948	< 1
Median late adaptivity	$+0.793$ (min $+0.227$)	–

Table 10: Interval width relative to plain conformal, as a median, a mean and a worst case, with the count of blow-ups, shown next to the marginal and hard-third coverage each scale buys with it. All arms scored on the same splits with the same plain-conformal reference.

Family	Scale	Median	Mean	Worst	$> 2\times$	$> 10\times$	Cov. / hard third
DreamerV3	Probe (ours)	$0.857\times$	$0.8\times$	$1.4\times$	0/40	0/40	0.906 / 0.766
	k-NN local	$0.953\times$	$1.3\times$	$4.3\times$	6/40	0/40	0.906 / 0.753
	Mondrian	$1.102\times$	$44.0\times$	$1011.8\times$	18/40	9/40	0.921 / 0.826
	Raw disagreement	$3.036\times$	$33.4\times$	$732.9\times$	23/40	11/40	0.910 / 0.760
TD-MPC2	Probe (ours)	$0.808\times$	$0.8\times$	$1.5\times$	0/19	0/19	0.902 / 0.785
	k-NN local	$1.018\times$	$2.3\times$	$7.2\times$	6/19	0/19	0.903 / 0.772
	Mondrian	$0.998\times$	$10.3\times$	$106.6\times$	3/19	2/19	0.922 / 0.804
	Raw disagreement	$1.178\times$	$20.1\times$	$219.1\times$	8/19	5/19	0.908 / 0.764

all 29 rows hold target coverage. The cost is sharpness, $0.948\times$ plain conformal on the compounding rows against the $0.66\times$ the headline reports against a stabilized ensemble; at $K=1$ the ensemble arm is degenerate by construction, so the comparator is plain conformal.

Width stability. Table 10 reports what each scale's hard-third coverage costs in width. The mean and the median disagree violently for Mondrian and for the raw disagreement scale: Mondrian sits at $1.10\times$ on the median DreamerV3 row but exceeds ten times the plain-conformal width on 9 of 40 rows, worst $1012\times$ on atari100k_jamesbond, and the raw scale behaves the same way, which is why the stabilized version exists. The probe never exceeds $1.4\times$ (DreamerV3) or $1.5\times$ (TD-MPC2) of the plain-conformal width on any row.

G Conditions and Limits

The probe is task-specific, and fitting it is cheap. Because the probe is fitted against measured error, two questions decide what it costs in practice. Neither has a comfortable answer alone, and together they bound the method usefully. A probe fitted on one task and carried unchanged to another keeps its coverage, on 104 of the 110 cross-task pairs among the 11 proprioceptive tasks, because the conformal quantile is still computed on the target. It keeps nothing else. Its late adaptivity on the target has median $+0.000$ against $+0.857$ for the target's own probe, with the two distributions not overlapping, and its intervals run a median $10.2\times$ wider. Ordering does not transfer, and validity does not need to (Figure 9). The price of refitting is small: coverage stays within ± 0.02 of target on all 62 rows at every training budget we tried, down to a single labeled episode, and eight labeled episodes recover a median 95% of the late adaptivity the full training split gives. The supervision this method needs is therefore per task and is roughly eight episodes of recorded reward, not a transferable artifact.

The same rule does not improve planning. A budget over a rollout is not yet a control decision, so we also tested the obvious use. A model-predictive planner scored each candidate by its imagined return up to the horizon the interval trusts, rather than over the full 40 steps, on 11 DMC tasks of which 7 passed a validity gate requiring the planner to reach at least half the trained actor's return. We pre-registered that probe truncation would beat no truncation on most tasks, and separately that it would beat both the same rule driven by ensemble disagreement and a fixed-horizon control. On the three tasks registered first, the criteria pass by sign count, 2 of 3 each, but the signs carry the whole

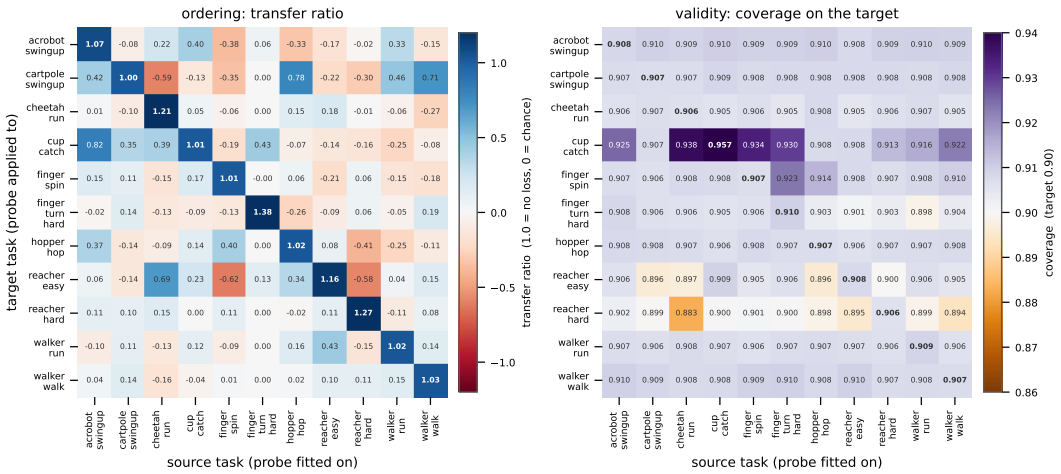

Figure 9: Cross-task transfer of the probe over all 11×11 source–target pairs of the DMC proprioceptive tasks. Left: the transfer ratio, the late-horizon adaptivity of a probe fitted on the source and applied to the target, divided by that of the target's own probe. The diagonal is intact (median 1.03) and the 110 off-diagonal cells have median 0.00, with only 4 above 0.5: a borrowed probe orders the target's rollouts no better than chance. Right: coverage of those same transferred intervals on the target, which stays within ± 0.02 of the 90% level on 104 of the 110 pairs. The two panels are the result in one picture. The conformal quantile is computed on target data, so validity does not depend on the probe being right, and what transfer costs is the ordering alone.

result: no task's paired interval excludes zero except cheetah_run, where truncation hurts by -7.0 return, and the largest gain is walker_walk at $+8.7$ $[-5.9, +23.0]$. On all 7 scored tasks, registered before these numbers were seen, the same criteria fail at 2, 3 and 3 of 7. We therefore report that the intervals do not improve planning, which matches the earlier null result for an uncertainty penalty.

We also tested the repair our own behavior-shift result prescribes. Recalibrating the rule on the planner's own rollouts, using eight open-loop episodes in which each plan is executed in full so the realized error is observed at every horizon, makes the accepted horizon sensible everywhere: 14.7 to 22.0 steps across five tasks, against 0.9 to 40.0 under the offline rule. It still does not improve return, failing both pre-registered criteria at 2 of 5, and the two arms whose intervals exclude zero are both losses. Calibration was the diagnosis, but calibration is not the fix.

We also removed replanning, on the reasoning that long-horizon error should matter most when a single plan is executed in full. It does not help either. Of 7 tasks only 3 pass the same validity gate, because open-loop control itself costs the planner a great deal: walker_walk falls to 50 against an actor at 176, reacher_easy to 82 against 173. On the 3 scored tasks truncation beats no truncation on none of them, at -0.2, -20.4 and -0.7 return, and no paired interval excludes zero. That the gate removes 4 of 7 tasks is itself worth recording, since a planner that loses half its return when replanning stops cannot report on an uncertainty rule.

The reason is visible in the accepted lengths. The rule is calibrated on evaluation rollouts and applied to the planner's own candidates, and the horizon it accepts does not survive that move: 2.5 steps on the evaluation rollouts of reacher_hard against 39.9 during planning, 13.1 against 37.6 on cheetah_run, and 6.1 against 0.9 on cup_catch. It shifts in both directions, so this is not a constant offset but the same exchangeability failure that a changed behavior policy produces (below). What the intervals support is deciding how far to trust a rollout, not how to act on it.

Joint multi-horizon bands. Per-horizon intervals do not bound a whole trajectory, and the probe does not help when they must. A faithful CF-RNN (Stankevičiūtė et al., 2021) needs $H/\alpha - 1 = 499$ calibration rollouts for its Bonferroni correction. With about 250 its quantile saturates and the band under-covers, down to 0.60 joint coverage. Among valid supremum-score bands, the probe scale is not narrower than a constant band, and our pre-registered criterion failed at 6 of 17 compounding tasks against a bar of 10. It is narrower than scaling by disagreement on 34 of 40. Shrinking it

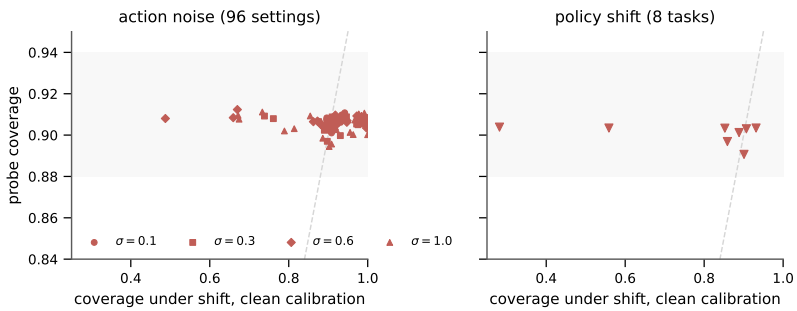

Figure 10: Behaviour-policy shift and the repair, on all 104 settings we ran. Each point is one setting: coverage with clean calibration under the shift (x) against coverage after recalibrating on a third of the new policy's episodes (y). Left: action noise on DreamerV3 at four levels (markers) and two collection arms. Right: the harsher TD-MPC2 policy shift, where a different trained member's planner collects the episodes and clean-calibrated coverage falls to 0.282 on cheetah_run. Recalibration lands every setting in the target band (shaded): 104 of 104, range 0.891–0.912; the dashed diagonal marks no change.

toward a constant, with the weight chosen on train episodes, reaches a median $0.91\times$ (DreamerV3) and $0.72\times$ (TD-MPC2) of the constant band at joint coverage 0.91.

Change of behavior policy. Conformal validity assumes calibration and test rollouts are exchangeable, and when that assumption fails the guarantee fails with it (Tibshirani et al., 2019). Under strong action noise, at levels 0.6 and 1.0 on the 11 proprioceptive tasks, intervals calibrated on the clean policy can under-cover, because the shifted latents leave the region the scale was fitted on. On cartpole_swingup probe coverage falls to 0.66 at noise 0.6 and 0.67 at noise 1.0, against 0.80 for plain conformal at both and 0.56 and 0.49 for ensemble scaling, and our pre-registered prediction that the probe would out-cover plain conformal there failed at 1 of 4 settings. Recalibrating on a third of the shifted policy's episodes restores probe coverage to 0.891–0.912 on all 104 settings we ran (Figure 10), which is the full action-noise grid over 4 levels and both collection arms together with the TD-MPC2 policy-shift rows, so the practical rule is to recalibrate when the behavior policy changes. Weighted conformal prediction (Tibshirani et al., 2019) corrects for a shift whose likelihood ratio is known. We do not have one, since the shift is induced by a change in the behavior policy we do not model, so we recalibrate on episodes from the new policy instead.

The same test on TD-MPC2 uses a different kind of shift: a different trained member's planner collects the episodes, rather than noise being added to one policy's actions. It is a harsher shift. Clean-calibrated probe coverage falls below 0.88 on 4 of 8 tasks and averages 0.772, reaching 0.282 on cheetah_run, while plain conformal averages 0.881, so a per-rollout scale fitted on one policy's latents is more exposed than a constant width that does not depend on them. Our pre-registered prediction that the probe would out-cover plain conformal failed here as it did on DreamerV3, at 2 of 8. Recalibrating on a third of the new policy's episodes restores probe coverage to 0.891–0.904 on all 8 tasks. Across both families, then, recalibration is the prescription that holds, and robustness without it is not something we claim.

Flat-error Atari. Where error does not compound, the probe is only as good as disagreement. At $H = 150$ on the four answerable Atari games, the `dis2` probe beats the stabilized ensemble on one, against a pre-registered prediction of most.

H Limitations in Full

The probe is linear, so the oracle ceiling we report bounds linear readouts only. That ceiling is only a ceiling where the in-sample fit is identified, which we take to be at least three training rows per feature. On the 34 tasks that qualify the probe reaches a median 87% of it. On the other 26, all of them Atari, DMC vision or Crafter with roughly as many features as rows, the oracle interpolates and reads 0.95 whatever the data, so we do not quote a fraction there. Nonlinear readouts of the same latent do not close the gap either. Three nonlinear readouts of the same latent do no

better: over the 20 tasks that need no dimensionality reduction, random Fourier features beat the linear probe on 1, a per-horizon probe on 5, and a quadratic expansion on 9, at a median adaptivity change of -0.001 (Appendix F). Readout capacity is therefore not the binding constraint. The headline tables use one ensemble per task, and their standard deviations are over episode splits rather than training seeds. On the three tasks we retrained, split resampling understates the seed spread of adaptivity by up to $4.3\times$, though the per-task verdicts reproduce. The growth-rate rather than the asymptotic form of Theorem 4 is what applies to these benchmarks (Section 4), and because the error does not grow on 21 of the 25 Atari games, the claim that the probe helps where error compounds rests on DMC and TD-MPC2. The intervals assume the calibration and test rollouts are exchangeable, so a strong change of behavior policy costs coverage until they are recalibrated, and they are per-horizon rather than simultaneous over a trajectory (Appendix G). Finally, they support the trusted-horizon decision but not every decision. Neither an uncertainty penalty nor trusted-horizon truncation improved real return, and a pre-registered safety filter built on the intervals passed on 2 of 13 tasks. Making the intervals useful inside a planner would need them calibrated on the planner's own rollout distribution, which is what our behavior-shift result prescribes and what we leave to future work.

I Baseline Specifications

We compare against five conformal variants and against the published uncertainty signals listed after them.

- **Plain conformal** (`conf`). $s(x) = 1$, giving constant width per horizon. It is valid and has no adaptivity.
- **Ensemble-scaled** (`ens`). $s(x) = \max(\text{disagreement}(x), \epsilon)$, the standard normalized conformal interval built on the unsupervised signal our theory expects to fail.
- **Mondrian** (`mond`). Bin by ensemble disagreement into 3 quantile bins, with one conformal quantile per bin. It uses the same signal as `ens` but asks only that the signal rank correctly, not that it have the right magnitude, which is the weaker demand.
- **k-NN local** (`knn`). Here $s(x)$ is the mean absolute error of the k nearest train rollouts in latent space, computed per horizon. It needs no fitted model and sees exactly the information our probe sees. We select $k \in \{5, 10, 20, 40\}$ on train episodes by the same width-minimizing rule that selects the probe's β.
- **CQR** (`cqr`). Conformalized quantile regression (Romano et al., 2019) fits a conditional τ-quantile of the error by pinball-loss gradient descent, then corrects it with an additive conformity score. The correction is additive rather than multiplicative, so a heavy-tailed prediction cannot blow the width up by $40\times$ the way the multiplicative scale does on pixel observations.
- **Oracle** (`oracle`). The probe fitted in-sample on the test episodes. This is not a method, since it sees the labels it is scored against. It is a ceiling on what a linear probe of these features can achieve, and we report it only when the fit is identified, meaning there are enough rows per feature.
- **Published uncertainty signals**, each used as the conformal scale. These are the MOReL-style maximum and the MOPO-style norm of ensemble disagreement, the latent-transition disagreement of Berger et al. (2026) in per-step and cumulative form, a last-layer Laplace approximation of the reward head with both Monte Carlo and linearized predictives, and the model's own predictive spread of reward. Where ensemble disagreement collapses to near zero, we also report a stabilized ensemble scale that uses the same train-selected floor as the probe. That stabilized version is the stronger comparator.

J Reproducibility Statement

The world-model configurations, the hyperparameter grids for the probe and the k-NN baseline, the conformal level, and the episode-split counts are in Appendix E; the baseline specifications are in Appendix I. The three-way episode split draws from `numpy.random.default_rng(0)`, and every run is deterministic given a fixed seed. Each pass-or-fail prediction was recorded with its bar in a log before the run that tested it, and the failures are reported in Section 6.