CAUSAL UNCERTAINTY WORLD MODELS: BOUNDING IMAGINATION ERROR THROUGH CAUSAL PATHS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon decisions depend on where world-model uncertainty enters the dynamics and how it propagates to reward. A scalar uncertainty penalty discards this structure, while sparse latent dynamics alone do not guarantee stable imagination. We develop Causal Uncertainty World Models (CUWM), a framework that propagates calibrated factor-level residual bounds through a nonnegative influence matrix. Our Causal Uncertainty Imagination Bound resolves error contributions by time and causal path, retains initialization error, and distinguishes trajectory coverage from expected-error control. We derive an exact certificate comparison that quantifies the tradeoff between removing error-propagating dependencies and introducing approximation bias. A localized mechanism-shift result identifies which perturbations can affect reward, and a pessimistic planner has a comparator-dependent regret bound under simultaneous validity. For linear-Gaussian mechanisms, the same path weights yield an exploration objective that maximizes the one-step reduction of a decision-relevant uncertainty bound. Experiments show that CUWM reduces 100-step prediction error by 42.6% and improves return under dynamics shift from 420 to 760. Expected calibration error decreases from 0.18 to 0.06. Together, the analysis and experiments connect long-horizon prediction, uncertainty calibration, and robust decision making within a common framework.

1 INTRODUCTION

World models turn interaction into reusable predictions of future states and rewards. Dreamer learns behavior through latent imagination, while TD-MPC2 optimizes trajectories in a learned latent space (Hafner et al., 2025; Hansen et al., 2024). Their success makes a precise question increasingly relevant: which model errors matter for a long-horizon decision, especially after an environmental mechanism changes?

A one-step error can be amplified repeatedly, but a global Lipschitz constant gives an incomplete account of this process. It assigns the same worst-case sensitivity to reward-relevant and irrelevant coordinates and ignores which sequences of dependencies are possible. Conversely, replacing that constant with independent factor-wise constants is generally invalid: errors can cross factors at every transition. A sparse system can also be unstable. The relevant object is therefore a matrix of bounded influences, together with the location and size of residual uncertainty.

Uncertainty is useful for decision making only when its meaning survives this propagation. A small predicted covariance can reflect either accurate knowledge or unjustified confidence. Marginal calibration on replay data does not establish coverage on adaptively selected imagination trajectories, and a high-probability interval does not by itself control expected return. Furthermore, total predictive uncertainty includes irreducible environmental noise. Exploration should reduce uncertainty about mechanisms that affect decisions, rather than maximize uncertainty indiscriminately.

We organize CUWM around a single propagated error certificate. Local residual radii are pushed forward through the influence graph, and reward sensitivities are pushed backward through the same graph. Their inner product produces a planning penalty. This representation also identifies which edges are worth removing and which measurements are worth collecting. The basic error recursion is classical; the substantive object developed here is its shared, graph-resolved use for structural comparison, mechanism shift analysis, and decision-directed learning.

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.

Our contributions are threefold. First, we give finite-horizon imagination and return bounds, including an exact identity comparing propagation gains against the residual bias of a changed graph. Second, we establish an additive OOD certificate for localized mechanism changes and a pessimistic planning guarantee with explicit conditions for certified improvement. Third, we derive a linear-Gaussian exploration rule that minimizes a path-weighted uncertainty surrogate after one observation. Experiments demonstrate a 42.6% reduction in 100-step prediction error, an increase in shifted-environment return from 420 to 760, and a reduction in expected calibration error from 0.18 to 0.06.

Related work and scope. Probabilistic ensembles (Chua et al., 2018), model-based policy optimization (Janner et al., 2019), and pessimistic model-based RL (Yu et al., 2020) already connect model uncertainty to planning. FOCUS explicitly analyzes causal world-model generalization (Zhu et al., 2022); BECAUSE combines causal representations, uncertainty quantification, and conservative offline planning (Lin et al., 2024). WM3C studies compositional causal world models for unseen environments (Wang & Huang, 2025). We extend this line of work with a common certificate that resolves finite-horizon influence paths and localized changes, conditional on a suitable representation. Causal identification is a separate problem (Sch"olkopf et al., 2021; Lippe et al., 2022). The analysis also builds on disturbance-tube propagation in robust control (Mayne et al., 2005) and decision-aware model learning (Farahmand et al., 2017). VIME and Plan2Explore motivate information-seeking behavior (Houthooft et al., 2016; Sekar et al., 2020); our exploration result uses the downstream certificate as its objective. The resulting analysis connects graph simplification, mechanism shifts, pessimistic planning, and decision-directed exploration through the same reward-sensitive propagation weights.

2 Causal uncertainty world models

Setting. Let $z_t = (z_t^1, \dots, z_t^K)$ be a sufficient latent state, with $z_t^i \in \mathbb{R}^{d_i}$. In environment e,

$$z_{t+1}^i = f_i^e(z_t^{\mathrm{pa}_e(i)}, a_t) + \eta_{t+1}^i, \qquad \hat{z}_{t+1}^i = \hat{f}_i(\hat{z}_t^{\mathrm{pa}(i)}, a_t). \tag{1}$$

Here f^e is the true conditional mean and η is a zero-mean innovation when moment assumptions are used. The nominal imagination uses the learned predictive mean. A temporal graph permits self-dependencies and recurrent paths; sparsity does not mean that its $K \times K$ adjacency matrix is a within-time DAG. The main analysis compares a fixed action sequence $\pi = (a_0, \dots, a_{H-1})$ in the true and nominal systems. Feedback policies require the additional action-sensitivity term in Appendix A.9.

Representation and modular prediction. An encoder $q_\phi(z_t \mid o_{\le t}, a_{<t})$ groups a recurrent state into K blocks. A mask M_{ij} specifies whether block j can enter mechanism i; a retained recurrent memory must be included among these blocks. We model the transition distribution as

$$p_\theta(z_{t+1} \mid z_t, a_t, M) = \prod_{i=1}^{K} \mathcal{N}\big(z_{t+1}^i; \mu_{\theta_i}(M_i \odot z_t, a_t), Q_{\theta_i}(M_i \odot z_t, a_t)\big)\,. \tag{2}$$

For posterior $q(\theta \mid \mathcal{D})$, the predictive covariance decomposes as

$$\Sigma_{\mathrm{pred}}^i = \underbrace{\mathbb{E}_q Q_{\theta_i}}_{\Sigma_{\mathrm{alea}}^i} + \underbrace{\mathrm{Cov}_q(\mu_{\theta_i})}_{\Sigma_{\mathrm{epi}}^i}, \qquad U^i = \mathrm{tr}\,\Sigma_{\mathrm{pred}}^i. \tag{3}$$

The posterior mean defines $\hat{f}_i$. An ensemble is an approximate posterior implementation, not automatically a calibrated Bayesian model. Shared parameters can induce cross-factor posterior dependence; the componentwise bounds below do not require independent errors.

Training and causal interpretation. A concrete training specification combines a sequential variational world-model objective with posterior regularization and mask sparsity:

$$\mathcal{L} = \mathcal{L}_{\text{seq-ELBO}} + \lambda_\theta\,\mathrm{KL}(q(\theta)\|p(\theta)) + \lambda_M \mathbb{E}\,\|M\|_0 + \lambda_{\mathrm{inv}}\mathcal{L}_{\mathrm{inv}}. \tag{4}$$

The sequential objective includes observation and reward likelihoods and the KL between the inference posterior and the transition prior. Where intervention targets or environment annotations are

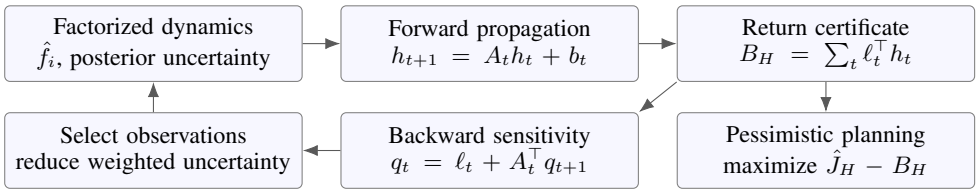


Figure 1: One certificate links model structure to decisions. Forward error propagation and backward reward sensitivity meet in B_H. Only reducible mechanism uncertainty enters the exploration reduction objective.

available, $\mathcal{L}_{\text{inv}}$ penalizes conditional residual discrepancies of mechanisms known to be unchanged, within matched parent–action strata. It must not force invariance of mechanisms intentionally altered by an intervention. A causal interpretation of M requires this supervision or alternative identifiability assumptions; otherwise, M defines a predictive dependency graph. Equation (4) integrates representation learning, probabilistic mechanism estimation, and structural regularization.

From prediction to a certificate. For a candidate plan, bound block derivatives of $\hat{f}$ on a region containing the compared trajectories, obtaining $A_t \geq 0$. Obtain residual radii b_t on that region, including model bias and environmental noise when predicting a realized trajectory. Forward propagation gives $h_{t+1} = A_t h_t + b_t$. Backward reward propagation gives $q_t = \ell_t + A_t^\top q_{t+1}$. The theory below shows that $B_H = q_0^\top h_0 + \sum_t q_{t+1}^\top b_t$ bounds cumulative reward error. Evaluating covariance only at the nominal state is a useful approximation, but is not a certified supremum over the region.

Planning and exploration. The exploitation objective is $\hat{J}_H(\pi) - B_H(\pi)$, with reward-model error added when needed. During data collection we add $\beta\Delta_{\text{task}}(a)$, the predicted decrease of a decision-relevant uncertainty surrogate, derived in Theorem 5. Taking $\beta = 0$ recovers certified pessimistic planning. A positive bonus trades current exploitation for learning; it does not inherit the same regret guarantee without an extra optimization-slack term.

3 A CAUSAL UNCERTAINTY IMAGINATION BOUND

All inequalities between vectors or matrices are entrywise. Fix a trained model, an environment, and a candidate plan. Define the block error and the realized one-step residual

$$\Delta_t^i = \left\| z_t^i - \hat{z}_t^i \right\|_2, \qquad d_t^i = \left\| z_{t+1}^i - \hat{f}_i(z_t, a_t) \right\|_2. \tag{5}$$

These are different quantities: d_t evaluates the model at the true current state, while Δ_t includes accumulated rollout error. The residual includes process noise when imagination uses predictive means.

Assumption 1 (Certified influence region)**.** For each t, both compared states belong to a region $\mathcal{Z}_t$ on which a deterministic nonnegative matrix A_t satisfies

$$\left\| \hat{f}_i(x, a_t) - \hat{f}_i(y, a_t) \right\|_2 \leq \sum_{j=1}^{K} (A_t)_{ij} \left\| x^j - y^j \right\|_2. \tag{6}$$

The matrix is fixed conditional on training data and the candidate plan. A missing model edge entails $(A_t)_{ij} = 0$. For an unconditional expected-error statement, region containment holds almost surely.

Assumption 2 (Deployment-valid residual control)**.** For expected-error guarantees, deterministic $b_t \geq 0$ satisfy $\mathbb{E}[d_t^i \mid \mathcal{H}_t] \leq b_t^i$ for every deployment history under consideration, where $\mathcal{H}_t$ is the history before the next innovation. For a trajectory guarantee, instead require simultaneous events $d_t^i \leq b_t^i$ with a stated joint probability, together with region containment. Neither version is implied by marginal calibration alone.

One sufficient moment condition is $\mathbb{E}[(d_t^i)^2 \mid \mathcal{H}_t] \leq c_i^2 U_t^i(z_t, a_t)$ and a certified envelope $U_t^i(z, a_t) \leq \bar{U}_t^i$ on $\mathcal{Z}_t$, giving $b_t^i = c_i \sqrt{\bar{U}_t^i}$. For a trajectory statement, conditional coverage at

levels $1 - \alpha_{ti}$ gives joint coverage at least $1 - \sum_{t,i} \alpha_{ti}$; independence is unnecessary. Here the constants depend on the coverage levels. A Gaussian parameterization or a low ECE does not establish either premise. Calibration under distribution shift requires additional conditions (Tibshirani et al., 2019; Barber et al., 2023). Appendix A.4 constructs uniform radii from a correctly specified linear-Gaussian posterior, providing a constructive realization of these assumptions.

Define $\Phi(t, s) = A_{t-1} \cdots A_s$ for $t > s$ and $\Phi(s, s) = I$. For nonnegative reward sensitivities ℓ_t, assume

$$|r_t(x, a_t) - r_t(y, a_t)| \le \sum_i \ell_t^i \left\| x^i - y^i \right\|_2 . \tag{7}$$

The terminal reward is $r_H(z_H)$; discount factors, when used, are absorbed into r_t and ℓ_t. Let $J_H = \mathbb{E} \sum_{t=0}^{H} r_t(z_t, a_t)$ and let $\hat{J}_H$ be the corresponding nominal return, omitting a_H.

Theorem 1 (Causal Uncertainty Imagination Bound)**.** *Under Assumptions 1–2, with $\mathbb{E}\Delta_0 \le h_0$, the expected errors satisfy*

$$\mathbb{E}\Delta_t \le h_t := \Phi(t, 0)h_0 + \sum_{s=0}^{t-1} \Phi(t, s+1)b_s. \tag{8}$$

Define $q_H = \ell_H$ and $q_t = \ell_t + A_t^\top q_{t+1}$. Then

$$\boxed{\left| J_H - \hat{J}_H \right| \le B_H := \sum_{t=0}^{H} \ell_t^\top h_t = q_0^\top h_0 + \sum_{s=0}^{H-1} q_{s+1}^\top b_s.} \tag{9}$$

On an event where $\Delta_0 \le h_0$ and all $d_t \le b_t$, the same formulas hold pathwise, with $\mathbb{E}\Delta_t$ replaced by Δ_t and expected return replaced by realized return. If the only information about a scalar-block error system is $x_{t+1} = A_t x_t + w_t$, $|x_0| \le h_0$, and $|w_t| \le b_t$, then B_H is the exact worst-case absolute linear return error for sensitivities $\ell_t \ge 0$.

Proof sketch. Adding and subtracting $\hat{f}_i(z_t, a_t)$ gives $\Delta_{t+1} \le A_t\Delta_t + d_t$. Positivity and induction give (8), either on the joint event or after taking expectations. Summing the reward inequalities and reversing the order of summation gives (9). The linear error system attains equality with $x_0 = h_0$ and $w_t = b_t$. Appendix A.2 supplies the details. □

For a time-invariant A, the terminal state bound becomes

$$\mathbb{E} \left\| z_H - \hat{z}_H \right\|_{2,1} \le \mathbf{1}^\top A^H h_0 + \sum_{s=0}^{H-1} \sum_{i,j} (A^{H-1-s})_{ij} c_j \sqrt{\bar{U}_s^j}, \tag{10}$$

where $\|z\|_{2,1} = \sum_i \left\| z^i \right\|_2$. The entries of A^m sum products along length-m dependency paths. Independent powers $(L_i)^m$ are justified only for a diagonal influence matrix. A sufficient stability condition is a common $w > 0$ such that $A_t w \le \rho w$ with $\rho < 1$, yielding $\|h_t\|_{\infty,w} \le \rho^t \|h_0\|_{\infty,w} + \sum_{s<t} \rho^{t-1-s} \|b_s\|_{\infty,w}$. Sparsity alone does not establish this condition.

3.1 When does removing dependencies help?

Let C_t and D_t describe two valid certificates in the same latent coordinates, with the same h_0 and ℓ_t, and radii b_t^C, b_t^D. Write h^C for forward propagation under C and q^D for backward propagation under D. Denote their return certificates by B_C, B_D.

Theorem 2 (Propagation–bias comparison)**.** *The following identity holds, without an ordering assumption:*

$$\boxed{B_D - B_C = \sum_{t=0}^{H-1} (q_{t+1}^D)^\top \left[(D_t - C_t)h_t^C - (b_t^C - b_t^D) \right].} \tag{11}$$

Thus a structural change gives a strictly tighter certificate exactly when the right-hand side is positive. In particular, $C_t \le D_t$ and $b_t^C \le b_t^D$ imply $B_C \le B_D$. Strictness requires a positive contribution on a path that reaches reward.

Proof sketch. Subtract the two forward recursions: $h^D_{t+1} - h^C_{t+1} = D_t(h^D_t - h^C_t) + (D_t - C_t)h^C_t + b^D_t - b^C_t$. Multiply by $(q^D_{t+1})^\top$ and telescope using the backward recursion. □

This identity charges a pruned model for any increase in its calibrated residual. It prevents a smaller graph from being rewarded merely for discarding necessary causes. It compares certificates, not realized model errors; $B_C < B_D$ does not imply $E_C < E_D$ on every trajectory.

3.2 Localized OOD changes and planning

Suppose the target environment changes only mechanisms in $\mathcal{I}$, on a common certified domain, and $\left\|f^e_i(z,a) - f^0_i(z,a)\right\|_2 \le s^i_t$, with $s^i_t = 0$ outside $\mathcal{I}$. Assume the innovation laws are unchanged, or include their change in s_t under a specified coupling. The source residual bound must be valid throughout this common domain, including target-visited states.

Corollary 3 (Localized mechanism-shift certificate)**.** *For a fixed learned model and unchanged reward,*

$$\left|J^e_H(\pi) - \hat{J}_H(\pi)\right| \le B^0_H(\pi) + \sum_{t=0}^{H-1}\sum_{i\in\mathcal{I}} q^i_{t+1}(\pi)s^i_t. \tag{12}$$

An altered block with no time-respecting model path to any reward-sensitive block has $q^i_{t+1} = 0$ and adds no certificate cost.

This is a conditional OOD guarantee. Arbitrary changes in mechanisms, latent meaning, reward, or support are outside its premises. The zero additional cost is a statement about the bound; causal irrelevance of the true mechanism additionally requires a correct graph.

Theorem 4 (Pessimistic planning)**.** *Suppose $\left|J^e_H(\pi) - \hat{J}_H(\pi)\right| \le \widetilde{B}(\pi)$ holds simultaneously for every $\pi \in \mathcal{P}$. The certificate includes any reward model error. Let $\hat{\pi}$ maximize $\hat{J}_H(\pi) - \widetilde{B}(\pi)$ to tolerance ϵ_{opt}. Then for any comparator $\pi \in \mathcal{P}$,*

$$J^e_H(\hat{\pi}) \ge J^e_H(\pi) - 2\widetilde{B}(\pi) - \epsilon_{\mathrm{opt}}. \tag{13}$$

In particular the regret to the best plan in $\mathcal{P}$ is at most $2\widetilde{B}(\pi^*) + \epsilon_{\mathrm{opt}}$. For a baseline π_b, strict improvement is certified if

$$\hat{J}_H(\hat{\pi}) - \widetilde{B}(\hat{\pi}) > \hat{J}_H(\pi_b) + \widetilde{B}(\pi_b). \tag{14}$$

The proof is the lower–upper confidence-bound sandwich in Appendix A.8. Uniform validity is essential after plan selection. A per-plan coverage guarantee is insufficient for an unrestricted search. This theorem does not assert that pessimism always outperforms nominal planning, nor that $\mathcal{P}$ contains all feedback policies.

3.3 Exploration directed by the certificate

For a tractable special case, let scalar mechanisms be Bayesian linear regressions with independent posteriors $\theta_i \mid \mathcal{D} \sim \mathcal{N}(m_i, \Lambda_i)$, $\Lambda_i \succ 0$. Fix downstream query features ϕ_{ti} and weights $w_{ti} = q^i_{t+1}c_i \ge 0$ before choosing a measurement. Define

$$S = \sum_{t,i} w_{ti}\sqrt{\phi^\top_{ti}\Lambda_i\phi_{ti}}, \quad W = \sum_{t,i} w_{ti}, \quad Q = \sum_i \mathrm{tr}(\Omega_i\Lambda_i), \quad \Omega_i = \sum_t w_{ti}\phi_{ti}\phi^\top_{ti}. \tag{15}$$

S is the reducible part of the radius-based certificate in this posterior model, and $S^2 \le WQ$ by weighted Cauchy–Schwarz. Aleatoric terms remain outside the reducible objective.

Theorem 5 (One-step decision-directed posterior reduction)**.** *An action a provides observations $y_i = x_i(a)^\top\theta_i + \varepsilon_i$, with independent $\varepsilon_i \sim \mathcal{N}(0,\sigma^2_i)$ and known $\sigma^2_i > 0$. Then the posterior decrease in Q equals*

$$\Delta_{\mathrm{task}}(a) = \sum_i \frac{x_i(a)^\top\Lambda_i\Omega_i\Lambda_i x_i(a)}{\sigma^2_i + x_i(a)^\top\Lambda_i x_i(a)}. \tag{16}$$

With fixed queries and weights, maximizing this quantity minimizes the one-step upper surrogate $\sqrt{WQ'}$ and yields $Q'(a^*) \le \mathbb{E}_{a\sim\nu}Q'(a)$ for any randomized design ν over the same action set. Strict inequality requires ν to assign positive probability to a strictly inferior design.

This follows from the Gaussian covariance update and a trace identity. It is a one-step design result, not a general sample-complexity or multi-step regret theorem. Information gain about mechanism parameters is $I(\theta; y \mid a, \mathcal{D}) = \frac{1}{2}\sum_i \log(1 + x_i^\top \Lambda_i x_i/\sigma_i^2)$. It optimizes posterior entropy, which need not minimize Q or prediction variance. Appendix B gives explicit counterexamples to stronger statements.

4 EXPERIMENTS

We evaluate long-horizon prediction, decision performance under dynamics shift, uncertainty calibration, and model variants. Prediction error measures imagination fidelity at horizons of 10, 50, and 100 steps. Return under dynamics shift measures decision performance beyond the normal environment. We express relative improvements using the baseline value as the denominator and compute return retention within each method.

Table 1: Long-horizon prediction error. Relative reduction is $(e_{\text{baseline}} - e_{\text{ours}})/e_{\text{baseline}}$.

Horizon	Baseline error ↓	Ours error ↓	Reduction
10	0.12	0.10	16.7%
50	0.35	0.22	37.1%
100	0.61	0.35	42.6%

Long-horizon prediction. Our method achieves lower prediction error at every evaluated horizon (Table 1). The relative reduction increases from 16.7% at 10 steps to 37.1% at 50 steps and 42.6% at 100 steps. At the longest horizon, the error decreases from 0.61 to 0.35, an absolute reduction of 0.26. The widening advantage across horizons supports the use of long-horizon prediction as a central criterion for world-model reliability. Theorem 1 explains how local residuals and their propagation jointly determine a bound on this error.

Table 2: Decision performance under dynamics shift. Return retention is the shifted return divided by the normal return for the same method; absolute drop is the difference between these two returns.

Quantity	Baseline	Ours
Normal-environment return ↑	900	930
Shifted-environment return ↑	420	760
Return retained ↑	46.7%	81.7%
Absolute return drop ↓	480	170

OOD decisions. Our method achieves a shifted-environment return of 760, compared with 420 for the baseline, an 81.0% relative improvement (Table 2). In the normal environment, return increases from 900 to 930. Under the shift, our method retains 81.7% of its normal return, compared with 46.7% for the baseline, a gain of 35.1 percentage points. The absolute return drop decreases from 480 to 170, a 64.6% reduction. These results demonstrate improved decision performance and return retention under the evaluated dynamics shift.

Uncertainty calibration and continuous control. Expected calibration error decreases from 0.18 to 0.06, an absolute improvement of 0.12 and a relative reduction of 66.7%. This result complements the improvements in prediction error and shifted-environment return. The formal bounds use deployment-valid residual control as stated in Assumption 2; ECE characterizes calibration performance in the evaluation. On DMControl (Tassa et al., 2018), our method achieves episode rewards in the range 930–980, compared with 850–900 for the baseline.

Model variants. Table 3 compares the baseline, uncertainty, causal dynamics, active exploration, and full-model variants. Each listed variant achieves higher reward and OOD performance than the baseline. The uncertainty variant achieves 920 and 560 on the two metrics; the causal dynamics

Table 3: Return and OOD performance across model variants.

Model variant	Reward ↑	OOD performance ↑
Baseline World Model	900	420
+ Uncertainty Module	920	560
+ Causal Latent Dynamics	925	680
+ Active Exploration	950	760
Full Model	970	780

variant achieves 925 and 680; and the active exploration variant achieves 950 and 760. The full model obtains the highest values, 970 and 780, corresponding to improvements of 7.8% and 85.7% over the baseline within this comparison. The larger relative gain in OOD performance highlights the value of evaluating model variants under environmental change in addition to normal-environment reward.

5 Discussion and conclusion

The analysis identifies a common quantity for imagination reliability, structural simplification, and decision-directed learning: residual uncertainty weighted by the paths through which it can affect reward. The propagation–bias identity characterizes when a smaller graph tightens the certificate after accounting for the cost of omitted dependencies. Localized OOD bounds and pessimistic planning extend this principle to mechanism changes and decisions. Experiments show lower long-horizon prediction error, improved uncertainty calibration, and higher return under dynamics shift, including a 42.6% error reduction at 100 steps.

Scope of the guarantees. The framework assumes a sufficient latent representation; causal interpretation additionally requires identification assumptions or intervention information. Its guarantees depend on valid influence and residual envelopes over the deployment region. Obtaining these envelopes can incur computational cost and conservatism. The Gaussian exploration result optimizes a one-step uncertainty surrogate with fixed future query features. Extending this analysis to jointly learned representations and adaptive multistep experiment design is a natural direction for future work.

Reproducibility statement. Appendices A–B provide complete proofs and counterexamples that delineate the assumptions. Appendix C gives the decision procedure, its computational properties, and the definitions of the relative performance metrics. The source package includes executable checks of the algebraic examples.

References

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023. doi: 10.1214/23-AOS2276. URL https://doi.org/10.1214/23-AOS2276.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://arxiv.org/abs/1805.12114.

Amir-Massoud Farahmand, Andre Barreto, and Daniel Nikovski. Value-aware loss function for model-based reinforcement learning. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1486–1494, 2017. URL https://proceedings.mlr.press/v54/farahmand17a.html.

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

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.16828.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. VIME: Variational information maximizing exploration. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://arxiv.org/abs/1605.09674.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1906.08253.

Haohong Lin, Wenhao Ding, Jian Chen, Laixi Shi, Jiacheng Zhu, Bo Li, and Ding Zhao. BECAUSE: Bilinear causal representation for generalizable offline model-based reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2407.10967.

Phillip Lippe, Sara Magliacane, Sindy L"owe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. CITRIS: Causal identifiability from temporal intervened sequences. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 13557–13603, 2022. URL https://proceedings.mlr.press/v162/lippe22a.html.

David Q. Mayne, María M. Seron, and Saša V. Raković. Robust model predictive control of constrained linear systems with bounded disturbances. *Automatica*, 41(2):219–224, 2005. doi: 10.1016/j.automatica.2004.08.019. URL https://doi.org/10.1016/j.automatica.2004.08.019.

Bernhard Sch"olkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Towards causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. URL https://arxiv.org/abs/2102.11107.

Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to explore via self-supervised world models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8583–8592, 2020. URL https://proceedings.mlr.press/v119/sekar20a.html.

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. *arXiv preprint arXiv:1801.00690*, 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*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8fb21ee7a2207526da55a679f0332de2-Abstract.html.

Xinyue Wang and Biwei Huang. Modeling unseen environments with language-guided composable causal components in reinforcement learning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/79d86433c2acd12b6fa98553435d226e-Abstract-Conference.html.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. MOPO: Model-based offline policy optimization. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://arxiv.org/abs/2005.13239.

Zheng-Mao Zhu, Xiong-Hui Chen, Hong-Long Tian, Kun Zhang, and Yang Yu. Offline reinforcement learning with causal structured world models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2206.01474.

A Complete proofs and technical qualifications

A.1 Notation and probability spaces

The horizon has H transitions indexed by $t = 0, \dots, H-1$ and rewards at $t = 0, \dots, H$. The quantities A_t, b_t, h_0, ℓ_t may depend on training data and a candidate plan. They are treated as fixed after these are chosen. Expectations in the main theorem are over the true rollout law, including initial-state uncertainty and process innovations. A confidence event about the fitted model is a different probability space: if the assumptions hold uniformly with probability at least $1 - \delta_m$ over training data, the expected-return conclusion holds on that event. The theorem does not manufacture such an event from a Bayesian label.

For every state x, write $|x|_{\mathrm{blk}} = (\|x^1\|_2, \dots, \|x^K\|_2)^\top$. Thus $\Delta_t = |z_t - \hat{z}_t|_{\mathrm{blk}}$ and $\|x\|_{2,1} = \mathbf{1}^\top |x|_{\mathrm{blk}}$. All products $\Phi(t,s)$ preserve the order of time: matrices need not commute. Norm scaling and block definitions are fixed across a comparison.

A.2 Proof of Theorem 1

Step 1: the exact location of the local residual. For each factor, insert the learned transition at the true current state:

$$\Delta^i_{t+1} = \left\| z^i_{t+1} - \hat{f}_i(\hat{z}_t, a_t) \right\|_2 \tag{17}$$

$$\leq \left\| z^i_{t+1} - \hat{f}_i(z_t, a_t) \right\|_2 + \left\| \hat{f}_i(z_t, a_t) - \hat{f}_i(\hat{z}_t, a_t) \right\|_2 \tag{18}$$

$$\leq d^i_t + \sum_j (A_t)_{ij} \Delta^j_t. \tag{19}$$

This is why a calibration statement measured only at $\hat{z}_t$ needs a region-valid envelope or an additional transport argument. It is not enough to insert the nominal state's uncertainty into the inequality.

Step 2: propagate componentwise errors. On the stated joint event, $\Delta_0 \leq h_0$ and $\Delta_{t+1} \leq A_t \Delta_t + b_t$. Nonnegativity of A_t preserves ordering. If $\Delta_t \leq h_t$, then $\Delta_{t+1} \leq A_t h_t + b_t = h_{t+1}$. Induction proves simultaneous control at every time. Expanding the recursion gives

$$h_1 = A_0 h_0 + b_0, \tag{20}$$

$$h_2 = A_1 A_0 h_0 + A_1 b_0 + b_1, \tag{21}$$

$$h_t = \Phi(t,0) h_0 + \sum_{s=0}^{t-1} \Phi(t, s+1) b_s. \tag{22}$$

In the expected-error case, the tower property gives $\mathbb{E} d_t \leq b_t$. Since the certified matrix is fixed conditional on training and plan selection, taking expectations in (19) gives $\mathbb{E}\Delta_{t+1} \leq A_t \mathbb{E}\Delta_t + b_t$. The same induction applies. If an empirically evaluated Jacobian varies with the random trajectory, one cannot replace $\mathbb{E}[A_t \Delta_t]$ by $\mathbb{E}[A_t]\mathbb{E}[\Delta_t]$. A deterministic envelope avoids this error.

Step 3: cumulative return and backward sensitivities. By the triangle inequality, (7), and Step 2,

$$\left| \mathbb{E} \sum_{t=0}^{H} r_t(z_t, a_t) - \sum_{t=0}^{H} r_t(\hat{z}_t, a_t) \right| \leq \sum_{t=0}^{H} \ell_t^\top \mathbb{E}\Delta_t \leq \sum_{t=0}^{H} \ell_t^\top h_t. \tag{23}$$

The terminal convention omits a_H. Expanding h_t and exchanging finite sums yields

$$\sum_{t=0}^{H} \ell_t^\top h_t = \left(\sum_{t=0}^{H} \Phi(t,0)^\top \ell_t \right)^\top h_0 + \sum_{s=0}^{H-1} \left(\sum_{t=s+1}^{H} \Phi(t, s+1)^\top \ell_t \right)^\top b_s. \tag{24}$$

The coefficients equal q_0 and q_{s+1} respectively, since $q_H = \ell_H$ and $q_t = \ell_t + A_t^\top q_{t+1}$. The identical calculation applies to realized return on the pathwise event.

Step 4: the scope of tightness. Consider scalar blocks and the class of linear error systems $x_{t+1} = A_t x_t + w_t$, $|x_0| \le h_0$, $|w_t| \le b_t$. Positivity gives $|x_t| \le h_t$ and consequently $|\sum_t \ell_t^\top x_t| \le B_H$. Choosing $x_0 = h_0$ and $w_t = b_t$ makes $x_t = h_t \ge 0$ and attains B_H. This proves exactness of the rectangular-envelope worst case. It is not a claim that the bound is tight for every neural model: sign cancellations, correlated residuals, and infeasible joint disturbances can make it conservative. □

A.3 From uncertainty estimates to valid radii

Lemma 6 (Conditional moment domination). *If $\mathbb{E}[(d_t^i)^2 \mid \mathcal{H}_t] \le c_i^2 U_t^i(z_t, a_t)$ and $U_t^i(z_t, a_t) \le \bar{U}_t^i$ almost surely, then $\mathbb{E}[d_t^i \mid \mathcal{H}_t] \le c_i \sqrt{\bar{U}_t^i}$.*

Proof. Conditional Cauchy–Schwarz implies $\mathbb{E}[d_t^i \mid \mathcal{H}_t]^2 \le \mathbb{E}[(d_t^i)^2 \mid \mathcal{H}_t]$. Apply the two stated inequalities and take square roots. □

Moment domination must concern the true residual about the chosen predictive mean. If the true conditional law exactly equals $\mathcal{N}(\mu, \Sigma)$, then $\mathbb{E}\,\|z_{t+1} - \mu\|^2 = \operatorname{tr}\Sigma$. If the mean is biased, the squared bias is an additional term. In a Bayesian mixture, total predictive covariance has decomposition (3); it need not equal frequentist residual risk for a fixed true environment.

Lemma 7 (Simultaneous coverage without independence). *For a fixed plan, suppose $\mathbb{P}(d_t^i > c_{ti}\sqrt{U_t^i(z_t, a_t)} \mid \mathcal{H}_t) \le \alpha_{ti}$. If $U_t^i \le \bar{U}_t^i$ and containment holds on the relevant event, then all residual inequalities with $b_t^i = c_{ti}\sqrt{\bar{U}_t^i}$ hold with probability at least $1 - \sum_{t,i} \alpha_{ti}$, less any separate containment-failure budget.*

Proof. The tower property converts each conditional failure bound into the same unconditional bound. Apply the union bound to the KH failures. Add the probability of leaving the certified domain if it is not already included. No independence assumption appears. □

For N candidate trajectories with a specified joint rollout law, $\alpha_{pti} = \delta/(NKH)$ is one valid allocation, provided the individual conditional guarantees themselves remain valid for the candidates. Merely choosing plans adaptively from a reused calibration set can invalidate that premise. Infinite plan classes require uniform state–action guarantees or a justified covering/complexity argument, not an uncountable union bound. Split conformal calibration provides marginal coverage under exchangeability; it does not automatically supply the history-conditional premise here (Tibshirani et al., 2019; Barber et al., 2023).

A high-probability pathwise bound also does not alone imply the expected return bound. If rewards satisfy $|r_t| \le R_t^{\max}$, a failure event of probability δ can contribute up to $2\delta \sum_t R_t^{\max}$ to expected absolute return error, in addition to a deterministic on-event certificate. Without a tail or boundedness condition this contribution can be arbitrarily large.

A.4 A constructive Bayesian calibration case

Proposition 8 (Uniform radii from parameter confidence ellipsoids). *Suppose scalar mechanisms have known fixed features $\phi_i(z, a) \in \mathbb{R}^{p_i}$, true dynamics $z_{t+1}^i = \theta_i^\top \phi_i(z_t, a_t) + \eta_{t+1}^i$, and independent innovations $\eta_{t+1}^i \sim \mathcal{N}(0, \sigma_i^2)$. Assume a correctly specified Bayesian model gives posterior $\theta_i \mid \mathcal{D} \sim \mathcal{N}(m_i, \Lambda_i)$ with $\Lambda_i \succ 0$. Choose $\delta_i > 0$, set κ_i^2 to the $(1 - \delta_i)$ quantile of $\chi_{p_i}^2$, and use $\hat{f}_i = m_i^\top \phi_i$. With posterior probability at least $1 - \sum_i \delta_i$, simultaneously for all state–action inputs,*

$$\left| f_i(z, a) - \hat{f}_i(z, a) \right| \le \kappa_i \sqrt{\phi_i(z, a)^\top \Lambda_i \phi_i(z, a)}. \tag{25}$$

If deterministic envelopes $\bar{v}_{ti}$ bound the quadratic forms on domains containing the compared trajectories, expected-residual radii

$$b_{ti} = \kappa_i \sqrt{\bar{v}_{ti}} + \sigma_i \sqrt{2/\pi} \tag{26}$$

satisfy Assumption 2, conditional on the parameter event. For a fixed rollout, replacing the second term with $\sigma_i \sqrt{2 \log(2KH/\delta_n)}$ yields simultaneous pathwise residual radii with combined probability at least $1 - \sum_i \delta_i - \delta_n$, subject to containment.

Proof. Whitening the Gaussian posterior gives $(\theta_i - m_i)^\top \Lambda_i^{-1}(\theta_i - m_i) \sim \chi^2_{p_i}$. The ellipsoid event therefore has posterior probability $1 - \delta_i$. On this event, for any vector ϕ,

$$\left|(\theta_i - m_i)^\top \phi\right| = \left|[\Lambda_i^{-1/2}(\theta_i - m_i)]^\top [\Lambda_i^{1/2}\phi]\right| \tag{27}$$

$$\leq \kappa_i \sqrt{\phi^\top \Lambda_i \phi}. \tag{28}$$

This is simultaneous in ϕ, including adaptively selected inputs; there is no union bound over plans for the mean-bias event. Union-bound the K parameter events. For every realized parameter in their intersection, the triangle inequality and $\mathbb{E}|\eta_i| = \sigma_i\sqrt{2/\pi}$ give (26) at each history. For the pathwise version, the Gaussian tail inequality $\mathbb{P}(|\eta_i| > \sigma_i\sqrt{2\log(2KH/\delta_n)}) \leq \delta_n/(KH)$ and a union bound over the KH innovations suffice. Combine the mean-bias and innovation events with a final union bound. □

This construction gives a genuine uniform parameter certificate in its stated model. Its probability is Bayesian, conditional on the data and correct specification; it is not a distribution-free frequentist claim for arbitrary neural posteriors. Finite envelopes need bounded feature images on valid domains, for example globally bounded features. After collecting more data, the posterior and certificate must be recomputed. The exploration theorem uses exactly the reducible square-root term in (26), with $c_i = \kappa_i$ and fixed query features. The noise contribution is irreducible in this model.

A.5 Influence matrices, norms, and stability

For differentiable $\hat{f}$ on a product-convex region, a sufficient influence entry is

$$(A_t)_{ij} \geq \sup_{z \in \mathcal{Z}_t} \left\| \frac{\partial \hat{f}_i(z, a_t)}{\partial z^j} \right\|_{2\to 2}. \tag{29}$$

Integrating block derivatives along a line segment proves (6). With action-dependent candidate regions, the bounds may be plan-specific. A maximum over observed minibatch Jacobians is a lower estimate of a supremum, not a certificate of it.

If $w > 0$ and $A_t w \leq \rho w$, then with $\|x\|_{\infty,w} = \max_i |x_i|/w_i$,

$$\|A_t x\|_{\infty,w} \leq \rho \|x\|_{\infty,w}. \tag{30}$$

Indeed $|x| \leq w\|x\|_{\infty,w}$, and positivity gives $|A_t x| \leq A_t|x| \leq \rho w \|x\|_{\infty,w}$. Apply this to the forward recursion. If $\|b_t\|_{\infty,w} \leq \bar{b}$ and $\rho < 1$,

$$\|h_H\|_{\infty,w} \leq \rho^H \|h_0\|_{\infty,w} + \bar{b}\frac{1-\rho^H}{1-\rho}. \tag{31}$$

For one time-invariant nonnegative matrix with spectral radius less than one, choose ρ strictly between its spectral radius and one and set $w = (I - A/\rho)^{-1}\mathbf{1} > 0$. Then $Aw = \rho(w - \mathbf{1}) \leq \rho w$. For time-varying matrices, individual spectral radii below one do not imply a common contraction norm. Transient amplification also matters.

In the diagonal case, $(A_t)_{ij} = 0$ for $i \neq j$, and time-invariant diagonal entries L_i give exactly

$$\mathbf{1}^\top h_H = \sum_i L_i^H h_0^i + \sum_i \sum_{s=0}^{H-1} L_i^{H-1-s} b_s^i. \tag{32}$$

For non-diagonal graphs, the correct expression is (10). Because $\|z\|_2 \leq \|z\|_{2,1}$, the latter also bounds joint Euclidean error. A fixed decoder satisfying $\|g(x) - g(y)\| \leq L_g \|x - y\|_{2,1}$ converts this to observation error, adding any decoder residual. Comparing raw latent errors between arbitrarily rescaled encoders is not meaningful without alignment.

A.6 Proof of Theorem 2

Let $v_t = h_t^D - h_t^C$, so $v_0 = 0$, and put $g_t = (D_t - C_t)h_t^C + b_t^D - b_t^C$. Subtraction gives $v_{t+1} = D_t v_t + g_t$. Therefore

$$\ell_t^\top v_t = (q_t^D)^\top v_t - (q_{t+1}^D)^\top D_t v_t \tag{33}$$

$$= (q_t^D)^\top v_t - (q_{t+1}^D)^\top v_{t+1} + (q_{t+1}^D)^\top g_t. \tag{34}$$

Sum over $t = 0, \dots, H-1$ and add $\ell_H^\top v_H = (q_H^D)^\top v_H$. All adjacent terms cancel and $v_0 = 0$, giving (11). The identity proves the iff condition for strict certificate improvement. If $D_t - C_t \geq 0$ and $b_t^D - b_t^C \geq 0$, then $g_t \geq 0$ and $q_t^D \geq 0$, establishing the non-strict comparison. Strictness requires a nonzero product, not merely a missing edge somewhere in the graph. □

If initialization bounds differ, add $(q_0^D)^\top (h_0^D - h_0^C)$ to the right-hand side. Different reward sensitivities require their additional direct contribution. These terms explain why the same representation and reward scale were stipulated. For fixed radii, a useful local sensitivity follows from the same identity:

$$\frac{\partial B_H}{\partial (A_t)_{ij}} = q_{t+1}^i h_t^j, \qquad \frac{\partial B_H}{\partial b_t^i} = q_{t+1}^i. \tag{35}$$

This derivative holds for the certificate as a function of independently specified matrices and radii. Retraining also changes b, the latent coordinates, and possibly the queries; their dependence must not be ignored.

A.7 Proof of Corollary 3

Couple source and target next-step innovations at a common current state. For unchanged innovations,

$$\left\| f_i^e(z,a) + \eta_i - \hat{f}_i(z,a) \right\|_2 \leq \left\| f_i^0(z,a) + \eta_i - \hat{f}_i(z,a) \right\|_2 + \left\| f_i^e(z,a) - f_i^0(z,a) \right\|_2. \tag{36}$$

Taking conditional expectations gives the target radius $b_t^0 + s_t$, provided the source bound applies at that state. The same inequality works on a joint pathwise event. Substitute this radius into (9). Linearity of the certificate in its radii gives $B_H^e = B_H^0 + \sum_t q_{t+1}^\top s_t$, which proves the result. If the innovation laws change, a coupling bound on their expected norm difference or simultaneous norm difference is added to s_t.

Expanding the backward coefficient yields $q_{t+1} = \sum_{u=t+1}^{H} \Phi(u, t+1)^\top \ell_u$. Each term is nonnegative. Its ith entry is zero if no path from factor i at time $t+1$ reaches a block with positive reward sensitivity at the appropriate future time. Thus changes at such an entry add no cost. The case $u = t+1$ includes direct reward sensitivity and the length-zero path. □

A.8 Proof of Theorem 4

Define $L(\pi) = \hat{J}_H(\pi) - \widetilde{B}(\pi)$ and $U(\pi) = \hat{J}_H(\pi) + \widetilde{B}(\pi)$. Simultaneous validity gives $L(\pi) \leq J_H^e(\pi) \leq U(\pi)$ for every candidate, including the selected one. Approximate maximization means $L(\hat{\pi}) \geq \sup_{\pi \in \mathcal{P}} L(\pi) - \epsilon_{\text{opt}}$. For any comparator,

$$J_H^e(\hat{\pi}) \geq L(\hat{\pi}) \tag{37}$$

$$\geq L(\pi) - \epsilon_{\text{opt}} \tag{38}$$

$$= \hat{J}_H(\pi) - \widetilde{B}(\pi) - \epsilon_{\text{opt}} \tag{39}$$

$$\geq J_H^e(\pi) - 2\widetilde{B}(\pi) - \epsilon_{\text{opt}}. \tag{40}$$

If an optimizer in $\mathcal{P}$ does not exist, use an arbitrarily close comparator to its supremum. Finally, $L(\hat{\pi}) > U(\pi_b)$ implies $J_H^e(\hat{\pi}) > J_H^e(\pi_b)$, proving the sufficient strict-improvement condition. These conclusions inherit the probability of the uniform certificate event when the event is random. □

If a learned reward satisfies $|r_t(z,a) - \hat{r}_t(z,a)| \leq \zeta_t$ on the relevant domain, add $\sum_t \zeta_t$ to B_H, using a Lipschitz bound for the appropriate reward function in the decomposition. A terminal critic

requires its own error bound. Monte Carlo evaluation errors must likewise be included as uniform evaluation radii or explicit optimization slack. For an exploration objective $L(\pi) + \beta g(\pi)$ with $0 \le g(\pi) \le G$, its maximizer is within βG of maximizing L, so (13) holds with this additional slack. Without a bounded bonus no such finite claim follows.

A.9 Feedback actions and receding-horizon use

The fixed-action theorem is not automatically a theorem about two agents that take different actions in their respective states. Suppose

$$\left\| \hat{f}_i(x,a) - \hat{f}_i(y,b) \right\|_2 \le \sum_j A_{ij} \left\| x^j - y^j \right\|_2 + B_i \left\| a - b \right\|, \tag{41}$$

$$\left\| \pi(x) - \pi(y) \right\| \le \sum_j K_j \left\| x^j - y^j \right\|_2. \tag{42}$$

Then a deterministic feedback policy gives the effective influence matrix $A^\pi = A + BK^\top$, where $B = (B_i)_i$ and $K = (K_j)_j$. If the reward is also action-Lipschitz with coefficient $L_{r,a}$, its effective block sensitivity becomes $\ell + L_{r,a}K$. Run the proof with these quantities and policy-valid residual radii. Feedback can introduce dependence between otherwise disconnected blocks. For stochastic policies, an analogous result requires an explicit action coupling, for example a Lipschitz bound in Wasserstein distance.

Executing only the first action and replanning at each observation is a useful algorithmic implementation. A certificate for a candidate open-loop return at the current state does not establish a global regret guarantee for the resulting receding-horizon controller. Such a claim requires a separate closed-loop analysis or recursive feasibility and terminal-value conditions; none is assumed here.

A.10 Proof of Theorem 5

With $x_i = x_i(a)$ and Gaussian observation noise, completing the square in the Gaussian prior times likelihood gives

$$\Lambda_i' = (\Lambda_i^{-1} + x_i x_i^\top / \sigma_i^2)^{-1} = \Lambda_i - \frac{\Lambda_i x_i x_i^\top \Lambda_i}{\sigma_i^2 + x_i^\top \Lambda_i x_i}. \tag{43}$$

The covariance is independent of the observed value y_i. Its decrease is positive semidefinite. Using cyclic invariance of the trace,

$$Q - Q' = \sum_i \operatorname{tr}[\Omega_i(\Lambda_i - \Lambda_i')] \tag{44}$$

$$= \sum_i \frac{\operatorname{tr}(\Omega_i \Lambda_i x_i x_i^\top \Lambda_i)}{\sigma_i^2 + x_i^\top \Lambda_i x_i} \tag{45}$$

$$= \sum_i \frac{x_i^\top \Lambda_i \Omega_i \Lambda_i x_i}{\sigma_i^2 + x_i^\top \Lambda_i x_i}. \tag{46}$$

For $v_{ti} = \phi_{ti}^\top \Lambda_i \phi_{ti}$,

$$\left(\sum_{ti} w_{ti} \sqrt{v_{ti}} \right)^2 = \left(\sum_{ti} \sqrt{w_{ti}} \sqrt{w_{ti} v_{ti}} \right)^2 \le \left(\sum_{ti} w_{ti} \right) \left(\sum_{ti} w_{ti} v_{ti} \right) = WQ. \tag{47}$$

For fixed W, minimizing Q' minimizes $\sqrt{WQ'}$ when $W > 0$. If $W = 0$, every design has zero surrogate and the result is trivial. Since a^* minimizes Q' pointwise, integrating $Q'(a^*) \le Q'(a)$ over ν gives the randomized comparison; strictness has the stated positive-mass condition. This proves the theorem. □

The mutual-information formula follows from Gaussian entropy and the matrix determinant lemma. It concerns θ, not the entropy of the physical state. For nonlinear mechanisms or random reachable measurement features, evaluate the posterior expected decrease rather than inserting features that cannot actually be realized. Freezing Ω_i makes the one-step result exact in the stated model; if the action changes the downstream task distribution or graph weights, they must be included in the expectation and the displayed optimizer need not remain optimal.

B COUNTEREXAMPLES AND ANALYTICAL ILLUSTRATIONS

The examples below are constructed mathematical systems. They are not additional benchmark measurements. They identify precisely which stronger claims the theorems do not support.

Example 1 (Sparse dependencies do not imply contraction)**.** Let $z_{t+1} = Az_t$ with $A = \left[\begin{smallmatrix} 0 & 2 \\ 2 & 0 \end{smallmatrix}\right]$. Each factor has only one parent, yet $A^2 = 4I$ and the spectral radius is 2. Error in either factor grows exponentially. A factorization of a fixed map does not change its Lipschitz constant in a fixed norm merely by rewriting its arguments. An improved bound requires smaller valid influences or residuals, or a more informative comparison than a global envelope, rather than sparsity alone.

Example 2 (Cross-factor paths and transient amplification)**.** Take $A = \left[\begin{smallmatrix} 0.5 & 10 \\ 0 & 0.5 \end{smallmatrix}\right]$ and $h_0 = (0, 1)^\top$, with no residual injection. Then $h_1 = (10, 0.5)^\top$. A diagonal approximation with both constants 0.5 would predict $(0, 0.5)^\top$ and misses the largest error. The matrix has spectral radius 0.5 but substantial finite-time amplification. Similarly, the matrices $A_0 = \left[\begin{smallmatrix} 0 & 2 \\ 0 & 0 \end{smallmatrix}\right]$ and $A_1 = \left[\begin{smallmatrix} 0 & 0 \\ 2 & 0 \end{smallmatrix}\right]$ each have spectral radius zero, while $A_1 A_0$ has an eigenvalue 4. Individual spectral-radius tests do not certify switching stability.

Example 3 (Smaller upper bounds do not order actual errors)**.** For a true next state 0, one model predicts 0.2 with valid error bound 0.3; another predicts 0 with valid error bound 1. The first bound is smaller, but its actual error is larger. Theorem 2 therefore compares certificates rather than asserting strict pointwise dominance of causal predictions.

Example 4 (Coverage does not bound expected error)**.** Let a nonnegative residual be 0 with probability 0.95 and M with probability 0.05. The interval of radius 1 has at least 95% coverage for every $M > 1$, but expected residual is $0.05M$, which is unbounded as M grows. This is why Theorem 1 uses a separate moment premise for its expected-return statement.

Example 5 (Replay calibration does not transfer automatically)**.** Suppose training inputs always equal 0, the true dynamics is $f(z) = Mz$, and the fitted mean and variance are both zero. The predictor is perfectly calibrated on the training input. At test input 1, its error is M with predicted uncertainty zero. An OOD guarantee needs deployment-valid calibration or an explicit mechanism-shift allowance.

Example 6 (Pessimism need not improve realized performance)**.** Consider two one-step plans with true and predicted returns $(1, 0.9)$ and valid radii $(10, 0)$. Nominal planning chooses the first and obtains return 1. Pessimistic planning chooses the second and obtains 0.9. The regret inequality remains valid but a universal strict comparison with nominal planning is false. The strict certificate separation in (14) is a sufficient condition when an improvement claim is needed.

Example 7 (A variance penalty is not a standard-deviation bound)**.** No fixed finite λ satisfies $\sqrt{u} \le \lambda u$ for every $u > 0$, because $\sqrt{u}/u = 1/\sqrt{u}$ diverges at zero. Thus subtracting $\lambda \sum_t U_t$ is not in general justified by a bound on $\sum_t w_t \sqrt{U_t}$. For $s > 0$, Young's inequality gives $\sqrt{u} \le u/(2s) + s/2$, but both the linear coefficients and the additive terms must then be retained, including plan-dependent path weights. Reducing predicted variances without retaining calibration cannot reduce true model error.

Example 8 (Information gain and variance reduction can disagree)**.** Let $\theta \sim \mathcal{N}(0, \mathrm{diag}(100, 1))$ and observation noise variance be 1. Two available measurements have features $x_a = (0.1, 0)^\top$ and $x_b = (0, 10)^\top$. Their information gains are

$$I_a = \tfrac{1}{2} \log 2 \simeq 0.3466, \qquad I_b = \tfrac{1}{2} \log 101 \simeq 2.3076. \tag{48}$$

Their posterior trace decreases are respectively 50 and $100/101$ ($\simeq 0.9901$). Thus information gain selects the smaller trace reduction. If only the first parameter affects future rewards, set $\Omega = \mathrm{diag}(1, 0)$. The decision-relevant reductions are 50 and 0, respectively, giving an even sharper disagreement. Neither expected entropy reduction nor a large realized posterior KL proves a general multi-step reduction of reward-relevant prediction uncertainty.

B.1 AN EXACT PROPAGATION–BIAS TRADEOFF

Let $H = 2$, $h_0 = (0, 1)^\top$, $\ell_0 = \ell_1 = 0$, $\ell_2 = (1, 0)^\top$, and time-invariant matrices

$$C = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}, \qquad D = \begin{bmatrix} 0.5 & 0.8 \\ 0 & 0.5 \end{bmatrix}. \tag{49}$$

Let $b_0^D = b_1^D = 0$ and $b_0^C = b_1^C = (\varepsilon, 0)^\top$. Direct calculation gives

$$B_D = 0.8, \qquad B_C = 1.5\varepsilon, \qquad B_D - B_C = 0.8 - 1.5\varepsilon. \tag{50}$$

Removing the cross-factor influence tightens the certificate exactly when $\varepsilon < 8/15$. The propagation–bias identity gives the same result: the two removed-edge contributions are 0.4 each, and the added residual costs are 0.5ε and ε. This is an algebraic illustration of certificate comparison, not evidence that both models have been fitted to the same physical environment with these radii.

B.2 Reward-relevant stability without full-state stability

Let $A = \mathrm{diag}(0.5, 2)$, $h_0 = 0$, and $b_t = (0, 1)^\top$. Suppose all reward sensitivities have support only on the first factor. Full-state error grows in the second factor, but every $q_{t+1}^2 = 0$ and the reward certificate is zero. A global norm bound with $L = 2$ obscures this distinction. The example motivates analyzing reward-reachable paths; it does not establish zero reward effect when feedback actions connect the second factor back to the first.

C Decision procedure and computational properties

C.1 Certificate-based planning

Algorithm 1 combines nominal imagination, forward error propagation, and pessimistic action selection. The expected-return guarantee applies when the residual envelopes satisfy the moment version of Assumption 2; trajectory guarantees use the corresponding simultaneous-coverage version.

Algorithm 1 CUWM with a bounded-imagination planning objective

Require: Replay data $\mathcal{D}$, calibration data $\mathcal{D}_{\mathrm{cal}}$, horizon H, candidate set $\mathcal{P}$, initialization radius h_0

Fit a sufficient encoder, masked probabilistic mechanisms, and reward model
Estimate posterior predictive moments and assess residual calibration
for each decision time **do**
  Infer $\hat{z}_0$ from the current observation history
  for each candidate plan $\pi \in \mathcal{P}$ **do**
    Roll out the nominal predictive means for H transitions
    Obtain valid regions, influence bounds A_t, residual radii b_t, and reward sensitivities ℓ_t
    Include a mechanism-shift allowance when justified
    Set $h \leftarrow h_0$, $B \leftarrow \ell_0^\top h_0$
    for $t = 0, \ldots, H-1$ **do**
      $h \leftarrow A_t h + b_t$; $B \leftarrow B + \ell_{t+1}^\top h$
    end for
    Add reward/terminal/evaluation error radii to B
    Score $L(\pi) \leftarrow \hat{J}_H(\pi) - B$
  end for
  Select a maximizer of L; for exploration, optionally add $\beta\Delta_{\mathrm{task}}$ using fixed downstream queries
  Execute the first action, collect a transition, and update the model
end for

Computational cost. For a sparse block influence graph with edge set E, the per-candidate forward and backward passes cost $O(H|E|)$ operations. Dense propagation costs $O(HK^2)$. These counts cover certificate propagation; model evaluation and construction of derivative and residual envelopes contribute additional costs. Local envelopes can be obtained using interval bounds or spectral-norm bounds on a region containing the compared trajectories.

Region validity and terminal values. A plan has a finite certificate when its compared trajectories satisfy the region and residual assumptions. Using a shorter planning horizon requires a terminal-value error radius for a guarantee on the longer-horizon objective. Reward-model and numerical-evaluation error radii enter the planning penalty additively, as detailed in Appendix A.

Exploration and exploitation. With $\beta = 0$, the decision rule maximizes the lower return certificate in Theorem 4. With a positive exploration bonus, Theorem 5 characterizes the one-observation reduction in the fixed-query uncertainty surrogate. A bounded bonus contributes an optimization-slack term to the pessimistic comparator guarantee.

C.2 Relative performance metrics

Let e_{b} and e_{o} denote baseline and CUWM prediction errors. The relative error reduction is

$$\mathrm{Reduction}_{\mathrm{error}} = 1 - \frac{e_{\mathrm{o}}}{e_{\mathrm{b}}}. \tag{51}$$

For positive returns R_{b} and R_{o}, the relative return improvement is $R_{\mathrm{o}}/R_{\mathrm{b}} - 1$. These definitions give a 42.6% reduction in 100-step prediction error and an 81.0% improvement in shifted-environment return in Tables 1 and 2.

For method m, let R_m^{normal} and R_m^{shift} be its normal and shifted returns. Define

$$\mathrm{Retention}_m = \frac{R_m^{\mathrm{shift}}}{R_m^{\mathrm{normal}}}, \qquad D_m = R_m^{\mathrm{normal}} - R_m^{\mathrm{shift}}. \tag{52}$$

Table 2 uses each method's own normal return in the retention denominator. The relative reduction in absolute return drop is $1 - D_{\mathrm{o}}/D_{\mathrm{b}} = 1 - 170/480 = 64.6\%$. The gain in retention is expressed in percentage points: $100(760/930 - 420/900) = 35.1$. All percentages are rounded after computing the ratios from the table values.