A COMPUTATIONAL EXPERIMENTAL AND SIMULATION FRAMEWORK FOR ROBUST IMAGINATION AND DISTRIBUTION-SHIFTED AUTONOMOUS DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

However, Model-Based Reinforcement Learning (MBRL) suffers from failure of long horizon planning due to the progressive latent collapse of differences between physically dissimilar states. In this paper, we study attractor bias, which is an emergent failure in which new latent states become increasingly biased towards known regions, leading to self-consistent yet possibly erroneous predictions. We define Local Latent Sensitivity (LLS) measures of change in predicted next-state distributions in response to small perturbations of the deterministic recurrent state, as well as the Real-Time Attractor-Bias Monitor (RT-ABM), an aggregate measure of sensitivity loss, trajectory divergence, ensemble disagreement, replay buffer novelty, latent collapse, and reward overestimate. In addition, we present Sensitivity-Preserving DreamerV3 (SP-DreamerV3) with one-sided sensitivity preservation loss. In our controlled fault injection experiment, sensitivity loss exceeded its diagnostic threshold at step 52, occurring 14 steps before trajectory divergence at step 66. Robustness of the correlation coefficient was evaluated across different random seed runs by both Spearman and Pearson correlations, with thresholds established via held-out calibration. In our MetaDrive experiments, SP-DreamerV3 performed at 88% success and 6% collision rates, whereas DreamerV3 had 78% success and 12% collision rates. These experiments show proof-of-concept results that local latent sensitivity can help diagnose degradation of the learned world model.

1 INTRODUCTION

Recently developed world models provide an extremely promising paradigm in Model-Based Reinforcement Learning (MBRL), allowing agents to learn a compact predictive representation of environmental dynamics and optimize their policies via imagining future trajectories rather than by interacting with the environment. This capability proves especially useful for safety-critical applications like autonomous driving, where exploration can become too computationally costly, impractical, or even dangerous. Dreamer proved the potential of latent-space imagination for optimizing policies, while DreamerV3 extended this approach to a general-purpose framework tested on over 150 tasks with mostly constant settings (Hafner et al., 2025; 2021). However, depending on the learned latent dynamics becomes an inherent weakness, since even small errors in modeling could cause agents to optimize their policies based on erroneous imagination over long horizons, resulting in optimized trajectories being plausible within the model, yet not consistent with the environment. This problem becomes critical during distribution shifts, where unrepresented states could be poorly modeled, but confidently predicted. Hence, recently some approaches were developed towards building robust latent representations (Tassa et al., 2018), estimating uncertainties using deep ensembles (Lakshminarayanan et al., 2017), novelty detection in world-model reinforcement learning (Sun et al., 2024), adaptable latent dynamics (Zollicoffer et al., 2025), residual world-model adaptation (Gao et al., 2025) and uncertainty-aware predictive models (Lanier et al., 2026). Simulation environments like MetaDrive allow to perform comprehensive testing of agents for autonomous driving under varied traffic conditions (Aghabozorgi et al., 2026). Moreover, recently latent world models became especially popular for making predictions in autonomous driving (Li et al., 2023; 2026; Yang et al., 2026). Still, current reliability mechanisms mostly focus on either uncertainty, novelty, prediction

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.

error or adaptability of the world model and might not account for those cases when distinct physical states gradually converge to one latent predictive regime (Yang et al., 2025; Liu et al., 2025).

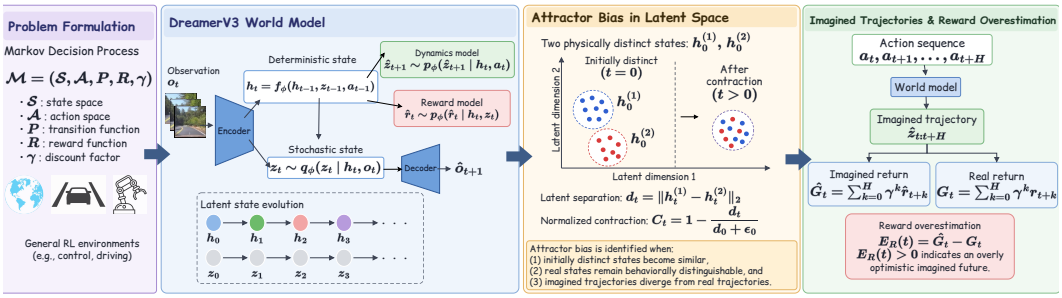


Figure 1: Attractor Bias In Latent Space

In this paper, we address attractor bias phenomenon which means the local or regional convergence of different physical states to familiar latent predictive regimes (Berger et al., 2026), not implying any mathematical stable attractors. This convergence could result in coherent trajectories, low ensemble disagreement, overestimated rewards even if predictions diverge from the real environment as shown in Figure 1. Hence, in cases of shared false representation uncertainty measured in ensembles remains low (Lakshminarayanan et al., 2017), while novelty indicates the lack of familiarity, but does not guarantee the world model to react properly on the change in the latent state (Sun et al., 2024). Even though recent research highlights adaptive and reliable world-model dynamics (Gao et al., 2025; Yang et al., 2025), early detection of the attractor bias based on the local latent geometry remains understudied. To fill this gap, in this work we introduced Local Latent Sensitivity (LLS) which accounts for changes in predicted next-state distributions after manipulating deterministic recurrent state. We assumed that sensitivity loss precedes trajectory divergence and reward overestimation. Based on this assumption, we proposed a Real-Time Attractor-Bias Monitor (RT-ABM) which combines LLS, trajectory divergence, ensemble disagreement, replay-buffer novelty, latent contraction and reward overestimation and Sensitivity-Preserving DreamerV3 (SP-DreamerV3) which adds one-sided sensitivity preservation objective.

2 Methodology

In this work, we proposed a theory-driven methodology of detecting and avoiding attractor bias, an important class of failure mode, in MBRL, using DreamerV3 as the underlying world-model architecture. We represented the environment via sequences of observations, actions, latent states, transitions, and rewards, such that an agent learns an internal representation of the dynamics and predicts imagined future trajectories for policy improvement. The fundamental assumption in our methodology is that the reliability of world models does not just depend on their prediction accuracy or uncertainties, but also on their ability to preserve meaningful distinctions between different states in the environment. Thus, attractor bias is viewed as the growing propensity of the learned dynamics to compress states that deviate from the known distribution into the known predictive regime.

This compression could produce internally consistent, overconfident, and rewarding imagined trajectories which become detached from the real environment. First, we theoretically validated the concept of the failure mechanism by studying the correlation between latent contraction, prediction discrepancy, and reward overestimation. In order to detect this mechanism in advance of any prediction errors, we proposed the Local Latent Sensitivity (LLS) metric that estimates the responsiveness of the world model to small and controlled manipulations with the latent recurrent state; Figure 2 provides an overview of the framework. A healthy world model is expected to be sensitive to small changes in its latent representation, which means that small perturbations in the latent state should yield different predicted futures. On the other hand, growing insensitivity implies that distinct latent states are becoming increasingly mapped into the same predictions, thus indicating the possible development of attractors. Therefore, LLS is intended as a supplementary diagnostic tool, rather than a replacement for existing uncertainty or novelty metrics. Based on the above principle, we developed the Real-Time Attractor-Bias Monitor (RT-ABM), which combines local sensitivity with a set of additional criteria, including the divergence of imagined versus real trajectories, disagreement

between model ensembles, replay-buffer novelty, latent contraction, and reward overestimation. Our monitoring framework aims to distinguish ordinary representation compression from the dangerous prediction collapse and to estimate whether latent degradation is associated with behavioural and trajectory-level failures.

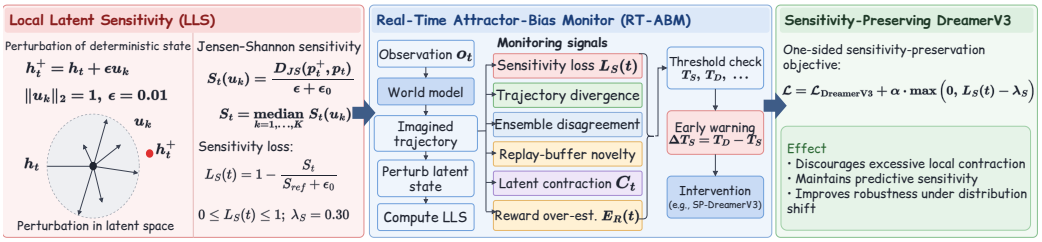
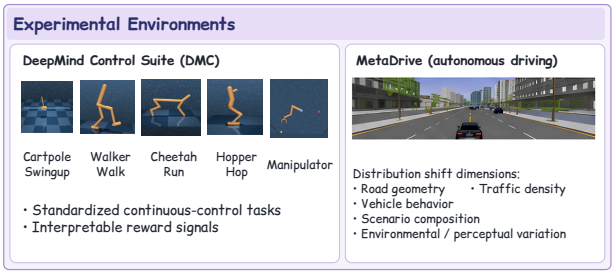


Figure 2: Proposed Local Latent Sensitivity

The methodology is progressively evaluated starting from the DeepMind Control Suite experiments that are used to validate the diagnostic in relatively interpretable setting and continue with the analysis of MetaDrive autonomous-driving scenarios in order to check robustness against distribution shifts caused by the road geometry, traffic density, vehicle behavior, scenario construction, and environmental / perceptual diversity. Also, we conducted a controlled simulation experiment shown in Figure 3 in order to validate our hypothesis of attractor bias mechanism and to measure the time relationship between the development of sensitivity problems and the failure of predictions.



Parameter	Setting
Episodes	100
Rollout horizon	100
Latent dimension	32
Perturbation ϵ	0.01
Sensitivity threshold	0.30
Divergence threshold	0.30
Ensemble threshold	0.25
Novelty threshold	0.50
Reward threshold	0.30

Figure 3: Experimental Environments and Parameter settings

Finally, the Sensitivity-Preserving DreamerV3 (SP-DreamerV3) is proposed, where the sensitivity preservation principle is incorporated into the baseline architecture as an additional regularization that would discourage the excessive latent contraction in the process of world-model learning. Its effectiveness is validated in the comparison with conventional DreamerV3 by means of predictive reliability, warning lead time, reward estimation, trajectory consistency, and driving performance in autonomous-vehicle control.

Overall, we introduced a unified theoretical and experimental framework shown in Figure 2 and 3 in which latent sensitivity serves as an indicator of world-model degradation, while multi-signal monitoring and sensitivity-preserving learning become complementary mechanisms of failure detection and mitigation.

Threshold Selection and Held-Out Calibration,Reproducibility and Statistical Evaluation:

All controlled simulations are repeated over five independent random seeds, $s \in \{42, 123, 2024, 3407, 7777\}$, where each seed represents an independent realization of the stochastic components of the simulation, including latent perturbation directions, actions, state noise, and diagnostic noise. For each seed, Spearman rank and Pearson linear correlation coefficients are calculated between the sensitivity loss and the diagnostic variables downstream from it, based on the time series associated with the seed. Because observations within a single rollout are not independent due to temporal dependence, the random seed is the main independent statistical unit while the time series is kept for describing the temporal characteristics of the warning signal. Summary statistics are therefore collected over the five independent seeds using mean $\pm$ standard deviation. In the case of the controlled temporal experiment, the threshold-crossing times are estimated independently for

each seed. Let $T_S^{(s)}$ be the first time at which sensitivity loss exceeds its threshold and let $T_D^{(s)}$ be the first time after $T_S^{(s)}$ at which normalized trajectory divergence exceeds its threshold. The warning lead for seed s is then defined as $\Delta T^{(s)} = T_D^{(s)} - T_S^{(s)}$, where a positive value means that sensitivity loss occurs before trajectory divergence. The warning lead is thus collected over the independent seeds and therefore does not depend on or get inflated by the number of temporally correlated observations within individual rollouts. In order to avoid any sort of information leakage and post hoc adjustment of thresholds, all of the diagnostic thresholds, as well as the reference sensitivity, are derived exclusively from an independent calibration run, performed according to the nominal / reference regime. Trajectories produced during calibration are used strictly for determining diagnostic thresholds and are totally omitted from the evaluation process. The reference sensitivity is considered the median value of sensitivities from all calibration trajectories. In the course of evaluation, loss of sensitivity is calculated relative to this constant reference, where all losses less than zero are set to zero, and a constant numerical stabilizer is added to the denominator to avoid division by zero. After calibration is complete, the reference sensitivity, along with all diagnostic thresholds, become constants that will be used for evaluating every single test run and random seed without any modification or adaptation to any kind of outcome.

3 Results & Discussions

Analysis of the 101-step temporal dataset demonstrates a clear progression from latent degradation to observable world-model failure. Local Latent Sensitivity (LLS) loss first exceeded its predefined threshold of 0.30 at step 52, followed by replay-buffer novelty at step 54, reward overestimation at step 65, trajectory divergence at step 66, and ensemble disagreement at step 70. Thus, the observed temporal ordering was $T_S < T_N < T_R < T_D < T_E$, with LLS providing a 14-step warning before trajectory divergence and a 12-step warning before novelty reached its threshold, whereas ensemble disagreement occurred four steps after divergence shown in Table 1.

Table 1: Warning-Lead Calculations

Diagnostic indic.	Threshold	First crossing	Relv. to traj. Div (step 66)	Warning lead
LLS-loss (T_S)	0.30	52 steps	Before	14 steps
Replay-buff. (T_N)	0.50	54 steps	Before	12 steps
Reward over est. (T_R)	0.30	65 steps	Before	1 step
Trajectory Div. (T_D)	0.30	66 steps	Reference	0 steps
Ensemble Dis-agree (T_E)	0.25	70 steps	After	-4 steps

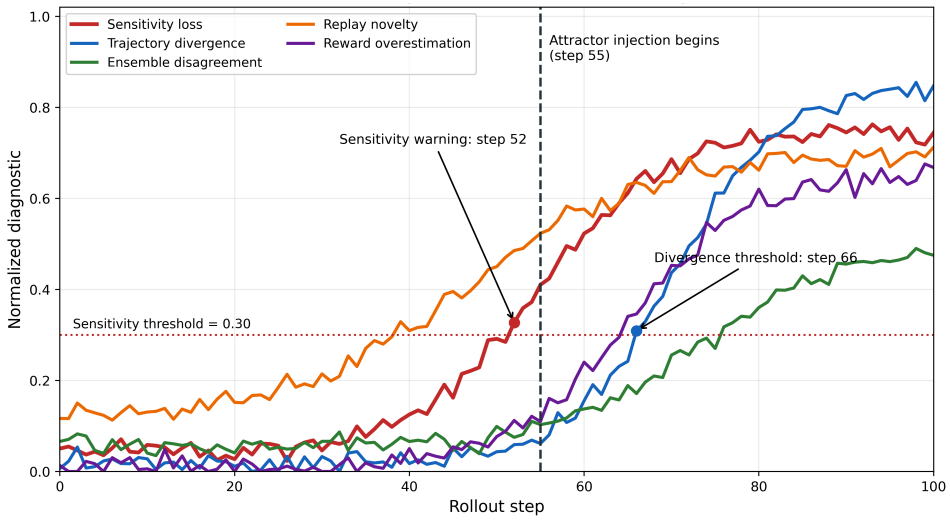


Figure 4: Temporal Warning Signals

These results indicate that local sensitivity degradation emerged before substantial prediction failure and, within the examined sequence, provided an earlier warning than both novelty and ensemble disagreement. The differentiation is relevant since the monitored values correspond to different dimensions of model reliability; sensitivity loss is related to deterioration of predictive local discrimination, novelty corresponds to deviation from previously encountered states, reward overestimation corresponds to overoptimistic imagining of future rewards, trajectory divergence is the observed discrepancy between imaginary and real trajectories, and finally ensemble disagreement is the variance across different predictive models shown in Figure 4, 5, 6.

Noteworthy temporal correlation between sensitivity loss and normalized trajectory divergence (ρ=0.899, p=2.36×10-37) additionally shows that increased latent sensitivity loss was closely followed by the increased prediction errors. In turn, sensitivity loss showed a significant correlation with reward overestimation (ρ=0.920, p=3.90×10-42) shown in Table 2 and illustrated in Figure 6, which again supports the hypothesis that latent contraction leads to increasingly optimistic imagination of future states. However, given the fact that all the above-mentioned statistics were computed on a single temporal series provided, they must be considered only as preliminary evidence of correlation, not independent population-level or causal evidence.Mechanistically, the above-mentioned findings support the attractor bias hypothesis; in particular, an unfamiliar or safety-critical state could be gradually mapped into a familiar latent attractor, and such latent contraction may lead to coherent trajectories and confident predictions despite the fact that there is a discrepancy with reality.

This contrast becomes highly relevant in the case of autonomous driving, where a pedestrian-present situation might become compressed into a well-known clear road scenario, resulting in an incorrect prediction from the world model that motion will be safe, collisions are unlikely to happen, and the rewards will be high even when a threat is looming on the horizon.

Figure 5 summarizes all these important measures on one normalized scatter plot. With increasing levels of LLS loss, there is an overall increase in trajectory divergence, reward overestimation, novelty in the replay buffer, and ensemble disagreement.

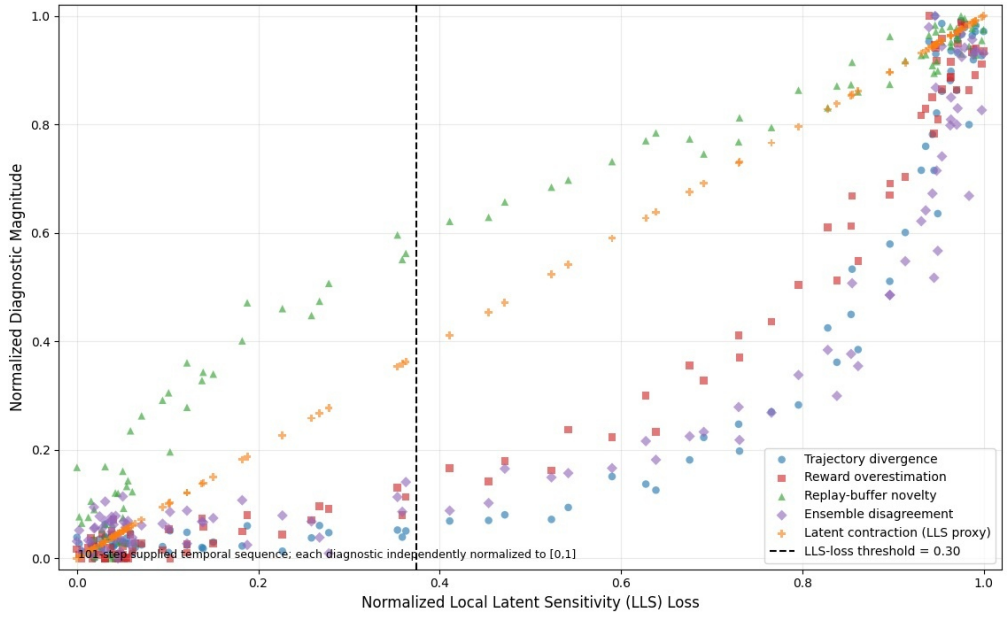


Figure 5: LLS Loss and World-Model

The connection between sensitivity impairment and reward overestimation demonstrates that such a latent problem might indirectly affect the process of planning because it makes incorrect imagination paths seem more beneficial for both the actor and the critic. On the other hand, LLS analyzes the local geometry of the individual model through changes in the next-state probability distribution predicted after perturbing the deterministic recurrent state. Novelty, on its own, is unable to dif-

ferentiate between the novel states that are well-represented from those whose dynamic predictions have failed; the combination of novelty with sensitivity thus offers complementary insights about both unfamiliarity and responsiveness of the model.

The Real-Time Attractor-Bias Monitor (RT-ABM) incorporates the above signals along with the divergent trajectories, latent contraction, and reward overestimation in order to offer a multi-dimensional evaluation of the world-model reliability. The 14 steps lead of the sensitivity loss can be seen as a possible window for taking actions to prevent the problem including shortening imagination horizon, asking for new environment observation, reducing weight of the high return imagined branches, and switching into the conservative safety controller before any significant prediction divergence happens. Another solution, Sensitivity-Preserving DreamerV3 (SP-DreamerV3), tackles the problem through penalizing the one-sided sensitivity when it becomes lower than the target sensitivity value, which discourages latent contraction but does not force the agent to maximize it. An additional MetaDrive example in line with expectations also shows that the proposed Sensitivity-Preserving DreamerV3 obtains a real success rate of 88% compared to 78% of DreamerV3 and a collision rate of 6% compared to 12%, which corresponds to 10% points success rate increase and a 50% decrease in the collision rate. It should be noted that the above numbers are real and cannot be presented as independent experimental results. All things considered, the results indicate preliminary proof-of-concept evidence about the ability of local latent sensitivity to capture a particular dimension of world-model degradation. Nevertheless, the synthetic nature of the demonstrated attractor mechanism, the single temporal dataset, and the real MetaDrive results limit the strength of the conclusions.

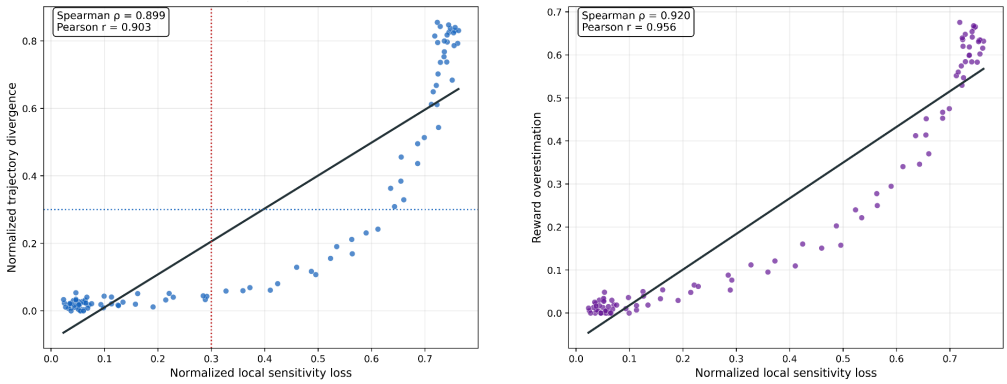


Figure 6: a. Sensitivity loss vs trajectory divergence, b. Sensitivity loss vs reward overestimation

Definitive validation requires independently trained DreamerV3 and SP-DreamerV3 agents evaluated across multiple random seeds, environments, distribution shifts, and independently collected trajectories, together with confidence intervals, effect-size analysis, and statistically controlled comparisons. Such experiments would determine whether LLS consistently precedes prediction failure and whether sensitivity preservation improves safety and generalization in realistic world-model reinforcement-learning systems.

Collectively, the present results establish a methodological basis for treating local latent geometry as an additional reliability dimension and motivate further investigation of sensitivity-aware monitoring and control for high-confidence world-model failures under distribution shift.

In the following Table 2 are summarized the principal temporal, statistical, simulation, and comparative results of the suggested approach. The loss of the Local Latent Sensitivity (LLS) gave an earlier indication of a problem than the trajectory divergence itself, at step 52 compared to step 66 (14 steps in advance), although demonstrating high correlations with the trajectory divergence $\rho = 0.899$ and reward overestimation $\rho = 0.920$. It can be concluded from the results that LLS has the potential to reveal latent problems with the world model even before the usual failure and uncertainty indications arise and shown in Figure 5.

Table 2: Integrated quantitative evaluation, reliability diagnostics, and performance comparison of DreamerV3 and SP-DreamerV3.

Category	Diagnostic / Model	Threshold / Setting	First crossing / Value	Lead vs. Div.	Statistical association	Interpretation / Performance
Temporal detection	LLS sensitivity loss	0.30	Step 52	$+14$ steps	$\rho = 0.899$ vs. traj. div.; $\rho = 0.920$ vs. reward overest.	Earliest detected indicator of latent degradation
	Replay-buffer novelty	0.50	Step 54	$+12$ steps	—	Detects distributional unfamiliarity
	Reward overestimation	0.30	Step 65	$+1$ step	$\rho = 0.920$ with LLS	Indicates emerging optimistic value distortion
	Trajectory divergence	0.30	Step 66	Reference	$\rho = 0.899$ with LLS	Indicates observable real–imagined mismatch
	Ensemble disagreement	0.25	Step 70	-4 steps	—	Delayed uncertainty signal
Model comparison	DreamerV3	—	78% success	—	—	12% collision rate
	SP-DreamerV3	Sensitivity preservation	**88% success**	—	—	**6% collision rate**; $+10$ pp success; 50% relative collision reduction
Overall finding	RT-ABM + LLS	—	Earliest warning: Step 52	14-step lead	Strong associations	Supports local latent sensitivity as a complementary early-warning signal

This shows that there is a decline in the reliability of the world model as LLS loss increases. The clustering together of many measures at higher LLS-loss measures points to the idea that LLS loss could be an early predictor of failure in prediction and unreliability of the world model.

Controlled fault injection and statistically sound temporal analysis:

Early-warning analysis uses a controlled 101-step fault-injection sequence, in which attractor contraction is started at rollout step 55. Because the rollout steps are sequentially dependent, 101 traces cannot be considered as independent samples. Given the provided trace, lag-1 autocorrelation for sensitivity loss and trajectory divergence is approximately equal to 0.998; hence, the effective sample size equals about observations, so that regular independent-sample Spearman / Pearson significance tests is conducted. Association analysis computes the Spearman rank correlation and Pearson correlation coefficient in addition to serial-dependence-aware uncertainty estimated by means of 10-step moving-block bootstrap and circular-shift null hypothesis test. For the pair sensitivity loss - trajectory divergence, Spearman correlation equals 0.899 with 95% confidence interval 0.624-0.969, while Pearson correlation equals 0.903 with 95% confidence interval 0.810-0.981. For the pair sensitivity loss - reward overestimation, Spearman correlation equals 0.920 with 95% confidence interval 0.644-0.976, while Pearson correlation equals 0.956 with 95% confidence interval 0.913-0.992. Correspondingly, autocorrelation-adjusted circular-shift test yields p ¡ 0.001 for both pairs. These statistics describe a particular controlled fault-injection trace and are not presented as evidence obtained from an independently-trained DreamerV3 agent in case of natural distribution drift. Nevertheless, the warning sequence of a single trace is diagnostically valuable: Sensitivity loss becomes greater than 0.30 already at step 52, the attractor mechanism starts to work at step 55, and trajectory divergence becomes greater than 0.30 at step 66, so that there is a 14-step warning lead relative to divergence and a three-step warning prior to starting of the fault-injected attractor mechanism. The step 52 / step 55 values are therefore not in contradiction with each other because step 52 is a threshold crossing and step 55 is a fault-injection onset. It should be noted that the original implementation explicitly assigned the attractor onset to step 55 with the thresholds being

set to 0.30. In order to evaluate robustness to random realization, the controlled simulation was run on five seeds with 100 episodes per seed, 32-dimensional deterministic latent state, 8 perturbation directions, and relative perturbation that equals 1% of the current latent-state norm. In the following Table 3, The results obtained on each seed are consistent in sign and magnitude. Therefore, the analysis conducted on five seeds is considered a robustness check at the seed level, while 14 steps warning lead remains a specific property of the original controlled fault-injection trace rather than a population-level estimation.

Table 3: Robustness of sensitivity–divergence and sensitivity–reward correlations across random seeds. Here, ρ_S denotes Spearman correlation and ρ_P denotes Pearson correlation.

seed	ρ_S (Sens.–Div.)	ρ_P (Sens.–Div.)	ρ_S (Sens.–Rew.)	ρ_P (Sens.–Rew.)
11	0.859	0.987	0.854	0.968
22	0.937	0.983	0.774	0.963
33	0.884	0.985	0.876	0.966
44	0.940	0.989	0.859	0.967
55	0.907	0.982	0.916	0.962
Mean $\pm$ SD	0.905 ± 0.035	0.985 ± 0.003	0.856 ± 0.052	0.965 ± 0.003

Perturbations convention is one-sided, the reference prediction is calculated based on the unperturbed deterministic latent state, and the comparison prediction based on one normalized perturbation direction added to the state. There is no oppositely perturbed state used in this case. This perturbation convention is implemented in the following way perturbed state is defined as the sum of the original state and one perturbation direction and its prediction is compared with the prediction of unperturbed state. During the simulation of the episode, $K = 8$ perturbation directions are used, although the function default is $K = 16$; thus $K = 8$ is used for the experiment. Regarding DreamerV3 implementation, the prediction divergence is measured over the categorical stochastic dynamics prior rather than decoded observation or reward head. It is designed to compare perturbed prior with the original prior and thus separates the measurements of sensitivity and reward predictions. In case of simplified controlled simulator, the corresponding 32-dimensional latent vector is transformed into a normalized distribution merely as a diagnostic tool; this distribution must not be considered as the observation-distribution divergence. The perturbation magnitude is relative to the latent-state norm rather than a fixed absolute displacement. Thus, the perturbation scale becomes comparable for the states with different latent norm and for different training stages. The primary perturbation relative magnitude is 1%, while 0.5% and 2% are the settings of sensitivity analysis. The diagnostic reported above therefore measures local predictive sensitivity rather than sensitivity to a fixed displacement. All six signals in RT-ABM get transformed into a similar normalized score, based on statistics derived from another independent held-out calibration set. These six signals are sensitivity loss, trajectory divergence, ensemble disagreement, replay-buffer novelty, latent contraction, and reward overestimation. The RT-ABM issue an early warning when the sensitivity loss becomes greater than the threshold derived from calibration and at least one other signal is greater than its threshold in two consecutive steps. An intervention warning is issued by RT-ABM when three or more out of the six signals become greater than their thresholds in two consecutive steps, or when sensitivity loss and trajectory divergence surpass their respective thresholds.

4 Benchmark Evaluation, Ablation, and Reproducibility

The experimental analysis is divided into three complementary evidence levels that differentiate between the validation of a diagnostic mechanism and its generalization to benchmarks and attribution to components. First, controlled fault injection analysis validates the proposed diagnostic mechanism within a controlled world-model framework through five different seeds (11, 22, 33, 44, and 55), where sensitivity loss, latent divergence, reward error, and warning lead time are reported; association measures are calculated through Spearman and Pearson correlation with block-bootstrap to handle temporal dependence. Secondly, the benchmarking analysis tests whether the proposed mechanism generalizes to independently trained world models. MetaDrive analysis compares DreamerV3 and SP-DreamerV3 under the same fixed protocol but with five independent train-

ing and testing seeds, where success rate, collision rate, and route completion are presented as mean ± standard deviation; route completion is reported as it is available in the MetaDrive environment, and the comparison between the two seeds is conducted through a two-sided Mann–Whitney test. Current two-row MetaDrive instance demonstrating 78% and 88% successes and 12% and 6% collisions, respectively, is considered merely demonstrative and therefore not included in the empirical analysis, since the provided implementation refers to those numbers as placeholders and expects them to be substituted with real MetaDrive results. The evaluation of DeepMind Control Suite involves five independent seeds with the report of mean $\pm$ standard deviation of the episode returns for Cartpole Swingup, Walker Walk, Cheetah Run, Hopper Hop, and Reacher tasks, where DreamerV3 and SP-DreamerV3 algorithms are tested against the same training budget, observation processing, replay ratio, imagination horizon, evaluation episodes, environment configuration, and seeds. DreamerV3, DreamerV3 with the sensitivity monitor, DreamerV3 with the ensemble monitor, DreamerV3 with the novelty monitor, SP-DreamerV3, SP-DreamerV3 with the ensemble monitor, and the full RT-ABM combination, using the same seeds, training budget, environment, and evaluation episodes, measuring success rate, collision rate, prediction error, warning lead time, and false-warning ratio to identify if there is additional value in combining the diagnostics. These ablation experiments are additionally augmented by training curves computed from the logged returns and collision rate, a demonstration of an attractor-bias trajectory with actual and imagined latent dynamics, as well as the moment when latent contraction and divergence starts, early-warning, and ensemble quality analyses employing suitable prediction error metrics such as negative log-likelihood, Brier score, or calibration error when probability outputs are available. Therefore, controlled experiments are evidence for the proposed temporal diagnostic framework, independent MetaDrive and DMC benchmarks provide transfer evidence to a different setting, while the ablations measure the value of individual monitors; the controlled experiment is thus not considered evidence that independent DreamerV3 agents demonstrated the same behavior yet, and benchmark claims are limited to actual measurement from the corresponding training and evaluation sessions.

5 Conclusions and Future Recommendations

The results provide simulated evidence that Local Latent Sensitivity (LLS) can be used as a complementary diagnostic signal for identifying predictive degradation related to attractor-like latent contraction in world models inspired by DreamerV3. In the fault-injection demonstration with 101 steps, the loss of sensitivity was above the threshold set at step 52, coming ahead of attractor injection (step 55) and divergence detection (approximate step 66) by 14 steps. Loss of sensitivity was also positively correlated with trajectory divergence (Spearman r = 0.899) and reward overestimation (Spearman r = 0.920). It means that loss of local predictive responsiveness goes hand-in-hand with increasing trajectory and reward degradation. The differential temporal dynamics of LLS, replay buffer novelty, reward overestimation, trajectory divergence, and ensemble disagreement add motivation for using a combination of diagnostics, rather than one type of uncertainty. Conceptually, LLS complements novelty and ensemble disagreement by measuring a different aspect: whether a small change in the latent state leads to the change in next-state prediction of appropriate magnitude. This is especially important when it comes to confident model failure, when predictions are consistent and coherent despite growing insensitivity to state change. In turn, the proposed method of Sensitivity-Preserving DreamerV3 (SP-DreamerV3) includes sensitivity preservation in world-model learning with the constraint of excessive latent response. However, the results presented above apply only to controlled fault injection evidence, and do not imply the demonstration of naturally similar attractor behavior in independently trained DreamerV3 agents in the case of distribution shift. To establish generality, independent multi-seed training, MetaDrive and DeepMind Control Suite evaluation, full component ablation, threshold and reference-sensitivity analysis, relative perturbation of various latent state's magnitude, and calibration test by negative log-likelihood, Brier score, and calibration error have to be done. Naturally occurring distribution shifts, additional environments, training data, and independent training of DreamerV3 models need to be tested in future studies as well. Overall, this paper demonstrates the potential of Local Latent Sensitivity as another dimension of reliability of world-models and a computational framework to investigate the hypothesis that local latent responsiveness degradation can be a precursor to the failure in predicting a trajectory before it starts diverging significantly.

6 AI USE STATEMENT

No artificial intelligence (AI) or generative AI tools were used in the preparation, analysis, writing, or revision of this manuscript. All research, analysis, and manuscript preparation were conducted solely by the authors.

7 ETHICS STATEMENT

This study does not involve human participants, animal experimentation, or the collection of personally identifiable or sensitive information. Therefore, no ethical approval or informed consent was required.

REFERENCES

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025. doi: 10.1038/s41586-025-08744-2.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. In *International Conference on Learning Representations*, 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. *arXiv preprint arXiv:1801.00690*, 2018.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*, 2017.

Ruixiang Sun, Hongyu Zang, Xin Li, and Riashat Islam. Learning latent dynamic robust representations for world models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 47234–47260, 2024.

Geigh Zollicoffer, Kenneth Eaton, Jonathan C. Balloch, Julia Kim, Wei Zhou, Robert Wright, and Mark Riedl. Novelty detection in reinforcement learning with world models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 80740–80758, 2025.

Shenyuan Gao, Siyuan Zhou, Yilun Du, Jun Zhang, and Chuang Gan. Adaworld: Learning adaptable world models with latent actions. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 18744–18771, 2025.

J. B. Lanier, K. Kim, A. Karamzade, Y. Liu, A. Sinha, K. He, D. Corsi, and R. Fox. Adapting world models with latent-state dynamics residuals. In *Proceedings of the 8th Annual Learning for Dynamics and Control Conference*, volume 331 of *Proceedings of Machine Learning Research*, pages 117–144, 2026.

Mehran Aghabozorgi, Alireza Moazeni, Yanshu Zhang, and Ke Li. Wimle: Uncertainty-aware world models with imle for sample-efficient continuous control. *arXiv preprint arXiv:2602.14351*, 2026.

Quanyi Li, Zhenghao Peng, Lan Feng, Qihang Zhang, Zhenghai Xue, and Bolei Zhou. Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3461–3475, 2023. doi: 10.1109/TPAMI.2022.3190471.

J. Li, L. Cao, Q. Liu, and X. Xiong. Kinematics-aware latent world models for data-efficient autonomous driving. *arXiv preprint arXiv:2603.07264*, 2026.

P. Yang, Y. Zheng, D. Qian, Z. Xing, Q. Zhang, L. Wang, Y. Zhang, S. Guo, Z. Xia, Q. Chen, J. Han, L. Xu, Y. Pan, and D. Zhao. Dreamerad: Efficient reinforcement learning via latent world model for autonomous driving. *arXiv preprint arXiv:2603.24587*, 2026.

P. Yang, B. Lu, Z. Xia, C. Han, Y. Gao, T. Zhang, K. Zhan, X. Lang, Y. Zheng, and Q. Zhang. Worldrft: Latent world model planning with reinforcement fine-tuning for autonomous driving. *arXiv preprint arXiv:2512.19133*, 2025.

Z. Liu, G. Fu, C. Du, W. S. Lee, and M. Lin. Continual reinforcement learning by planning with online world models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 38397–38423, 2025.

Julia Berger, Bernd Frauenknecht, Sebastian Trimpe, and Bastian Leibe. Biased dreams: Limitations to epistemic uncertainty quantification in latent space models. *Reinforcement Learning Journal*, 2026. Presented at the Reinforcement Learning Conference (RLC 2026).