Biased Dreams: Limitations to Epistemic Uncertainty Quantification in Latent Dynamics Models

Julia Berger, Bernd Frauenknecht, Sebastian Trimpe, Bastian Leibe

Keywords: MBRL, latent dynamics models, epistemic uncertainty, RSSM

Summary

Model-based reinforcement learning distinguishes between dynamics models operating on proprioceptive states and latent dynamics models typically operating on high-dimensional image observations. Among the latter, Dreamer's Recurrent State Space Model (RSSM) has emerged as a dominant architecture. While ensemble-based epistemic uncertainty has proven effective in proprioceptive dynamics for mitigating model exploitation, guiding exploration, or promoting caution, its behavior in latent dynamics remains largely unexplored. Our experiments reveal that although ensemble disagreement captures *local* epistemic uncertainty, it does not reliably reflect *global* compounding model error accumulated over prolonged RSSM latent rollouts. We provide evidence for an attractor behavior that draws rollouts toward well-supported latent regions, where uncertainty diminishes despite increasing discrepancies from the true environment dynamics. This can cause the model to overestimate returns when attractor regions correspond to high-reward behaviors. Our findings reveal a structural limitation of epistemic uncertainty estimation in RSSMs and challenge the assumption that epistemic uncertainty estimation transfers directly from proprioceptive to latent dynamics.

Contribution(s)

1. *Latent Attractor Behavior.* We provide empirical evidence that Recurrent State Space Models (RSSMs) can exhibit an attractor behavior, where latent transitions are biased toward well-supported latent regions that may diverge from the true environment dynamics.
   Context: Proprioceptive dynamics models typically accumulate prediction errors during prolonged rollouts, leading to increasingly unstable predictions.
2. *Limits of Epistemic Uncertainty.* Although ensemble disagreement reliably captures *local* epistemic uncertainty, we demonstrate that the attractor behavior can decouple uncertainty estimates from *global* trajectory-level model error in RSSM latent rollouts.
   Context: Ensemble-based epistemic uncertainty quantification is well-established for proprioceptive dynamics models and is often transferred directly to latent dynamics models without a careful analysis of its reliability.
3. *Return Overestimation in Latent Rollouts.* We observe that the attractor behavior can result in overly optimistic return predictions in RSSM latent rollouts when attractor regions coincide with high-reward behaviors.
   Context: This high-reward bias during imagination may distort policy learning by reinforcing overconfident behaviors that do not correspond to true environment outcomes.

Biased Dreams: Limitations to Epistemic Uncertainty Quantification in Latent Dynamics Models

Julia Berger[1, *], **Bernd Frauenknecht**[1, *], **Sebastian Trimpe**[1], **Bastian Leibe**[1]
{berger,leibe}@vision.rwth-aachen.de
{bernd.frauenknecht,sebastian.trimpe}@dsme.rwth-aachen.de

[1]**RWTH Aachen University**

[*] Equal contribution

Abstract

Model-based reinforcement learning distinguishes between dynamics models operating on proprioceptive states and latent dynamics models typically operating on high-dimensional image observations. Among the latter, Dreamer's Recurrent State Space Model (RSSM) has emerged as a dominant architecture. While ensemble-based epistemic uncertainty has proven effective in proprioceptive dynamics for mitigating model exploitation, guiding exploration, or promoting caution, its behavior in latent dynamics remains largely unexplored. Our experiments reveal that although ensemble disagreement captures *local* epistemic uncertainty, it does not reliably reflect *global* compounding model error accumulated over prolonged RSSM latent rollouts. We provide evidence for an attractor behavior that draws rollouts toward well-supported latent regions, where uncertainty diminishes despite increasing discrepancies from the true environment dynamics. This can cause the model to overestimate returns when attractor regions correspond to high-reward behaviors. Our findings reveal a structural limitation of epistemic uncertainty estimation in RSSMs and challenge the assumption that epistemic uncertainty estimation transfers directly from proprioceptive to latent dynamics.

1 Introduction

Reinforcement Learning (RL) provides a powerful framework for solving complex control problems (Berner et al., 2019; Degrave et al., 2022). Model-Based Reinforcement Learning (MBRL) addresses the sample inefficiency of standard RL by learning a parametric model of the environment dynamics to generate artificial interaction data for policy learning. Within the MBRL literature, we distinguish between *proprioceptive dynamics models* (Chua et al., 2018), which operate on proprioceptive inputs, and *latent dynamics models*, which typically learn compact state representations from high-dimensional image observations (Finn & Levine, 2017; Ha & Schmidhuber, 2018). Among latent approaches, Dreamer (Hafner et al., 2019a; 2020; 2023) has established the Recurrent State Space Model (RSSM) (Hafner et al., 2019b) as a dominant architecture for vision-based MBRL. While alternative latent dynamics have been proposed (Hansen et al., 2024; Zhou et al., 2024; Assran et al., 2025), RSSMs remain a widely employed and computationally efficient baseline.

A central question in MBRL is whether model-based rollouts provide reliable predictions for decision-making, as compounding model error can deteriorate policy learning (Janner et al., 2019; Yu et al., 2020). Ensemble-based epistemic uncertainty quantification is widely leveraged to identify unreliable model predictions (Lakshminarayanan et al., 2017). While this *local* epistemic uncertainty has been shown to be informative about *global* trajectory-level model error accumulated during rollouts in proprioceptive dynamics models (Janner et al., 2019; Yu et al., 2020; Frauenknecht et al., 2024), this relationship is often implicitly assumed to transfer to latent dynamics (Wang et al.,

2020; Zhu et al., 2020; Seyde et al., 2020; Filos et al., 2022; Seo et al., 2025; Liu et al., 2026). In this work, we study whether this relationship transfers to RSSMs and provide, to the best of our knowledge, the first systematic analysis of epistemic uncertainty quantification in this setting. We summarize our findings below, conceptually illustrated in Fig. 1:

(i) RSSM latent rollouts can exhibit an attractor behavior, where transitions are biased toward well-supported regions in latent space that may diverge from the true environment dynamics.

(ii) The attractor behavior can mask discrepancies between latent and environment dynamics, causing epistemic uncertainty estimates to become decoupled from compounding model error.

(iii) Attractor regions may coincide with high-reward behaviors, causing RSSM latent rollouts to predict overly optimistic returns.

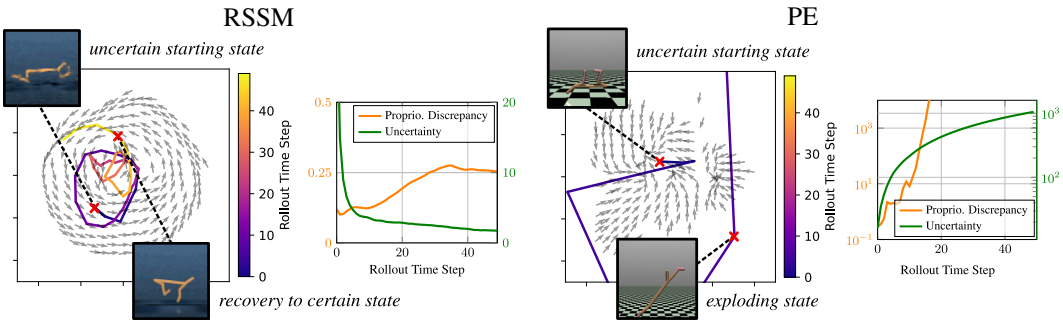

Figure 1: Conceptual illustration of our key findings. For an RSSM and a proprioceptive probabilistic ensemble (PE) model, the left subplots show PCA-embedded transition dynamics with an exemplary rollout initialized from an out-of-distribution state. The right subplots show proprioceptive discrepancy, which is the mismatch between predicted and true proprioceptive states under the same action sequence, and ensemble-based epistemic uncertainty. While the PE rollout becomes unstable with increasing prediction error and high uncertainty, the RSSM rollout converges to familiar latent regions, where uncertainty diminishes despite elevated proprioceptive discrepancy.

2 Background

In the following, we introduce the foundations of Reinforcement Learning (RL), Recurrent State Space Models (RSSMs), and uncertainty quantification.

2.1 Reinforcement Learning under Full and Partial Observability

Reinforcement Learning (RL) addresses sequential decision-making via agent–environment interaction. For an observable proprioceptive environment state $\mathbf{s}_t \in \mathcal{S}$, the control problem can be formulated as a Markov Decision Process (MDP) (Barto & Sutton, 2018). At each step, the agent selects an action $\mathbf{a}_t \sim \pi(\cdot \mid \mathbf{s}_t)$, after which the environment provides the next state and reward $\mathbf{s}_{t+1}, r_t \sim p(\cdot, \cdot \mid \mathbf{s}_t, \mathbf{a}_t)$. The objective is to learn a policy maximizing the expected discounted return $\mathbb{E}_\pi \left(\sum_{t=1}^{\infty} \gamma^t r_t\right)$ with discount factor $0 < \gamma < 1$. In Model-Based Reinforcement Learning (MBRL), a learned model $\tilde{p}(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)$ approximates the unknown environment dynamics and enables planning or simulated interactions. Early approaches employ Gaussian processes (Deisenroth & Rasmussen, 2011), local linear models (Levine & Koltun, 2013), and deterministic neural networks (Williams et al., 2017; Nagabandi et al., 2018). Recent successful MBRL methods on proprioceptive inputs (Chua et al., 2018; Janner et al., 2019; Yu et al., 2020) typically choose the probabilistic ensemble (PE) (Lakshminarayanan et al., 2017) model architecture.

In visual control settings, the underlying state $\mathbf{s}_t$ is not directly observable, and the problem is therefore formulated as a Partially Observable Markov Decision Process (POMDP) (Barto & Sutton, 2018). At each step, the agent selects an action $\mathbf{a}_t \sim \pi(\cdot \mid \mathbf{o}_{\leq t}, \mathbf{a}_{<t})$ conditioned on past

observations and actions, after which the environment provides the next observation and reward $\mathbf{o}_{t+1}, r_{t+1} \sim p(\cdot, \cdot \mid \mathbf{o}_{<t}, \mathbf{a}_{<t})$. The agent typically maintains a belief state summarizing the history of observations and actions used in decision-making. Because modeling dynamics directly in high-dimensional observation space is impractical, many approaches learn a compact latent representation that summarizes this history and enables predictive transitions (Finn & Levine, 2017; Ha & Schmidhuber, 2018). Such models are commonly referred to as latent dynamics models.

2.2 Latent Dynamics Modeling via Recurrent State Space Models

The Recurrent State Space Model (RSSM) (Hafner et al., 2019b) is a latent dynamics model that represents a belief state using stochastic state $\mathbf{z}_t$ and deterministic recurrent state $\mathbf{h}_t$. It consists of

$$\text{Representation model:} \quad q(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1}, \mathbf{o}_t), \tag{1}$$

$$\text{Observation model:} \quad p(\mathbf{o}_t \mid \mathbf{z}_t), \tag{2}$$

$$\text{Transition model:} \quad p(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1}), \tag{3}$$

$$\text{Reward model:} \quad p(r_t \mid \mathbf{z}_t). \tag{4}$$

For simplicity, we omit explicit dependencies $\mathbf{z}_{<t-1}, \mathbf{a}_{<t-1}$ induced by the recurrence of $\mathbf{h}_t$. All distributions are modeled as multivariate Gaussians with diagonal covariance.

We define two types of latent rollouts starting at some initial state $\mathbf{z}_0$. *Prior* rollouts τ^{prior} are generated solely by the transition model (Eq. (3)) and *posterior-informed* rollouts $\tau^{\text{post-inf}}$ perform one-step transition model predictions (Eq. (3)) initialized from posterior latent states $\mathbf{z}_{t-1} \sim q(\mathbf{z}_{t-1} \mid \cdot)$:

$$\tau^{\text{prior}} = \{(\mathbf{z}_t, \mathbf{a}_t)\}_{t=0}^{T-1}, \quad \mathbf{z}_t \sim p(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1}); \tag{5}$$

$$\tau^{\text{post-inf}} = \{(\mathbf{z}_t, \mathbf{a}_t)\}_{t=0}^{T-1}, \quad \mathbf{z}_t \sim p(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1}) \quad \text{and} \quad \mathbf{z}_{t-1} \sim q(\mathbf{z}_{t-1} \mid \cdot). \tag{6}$$

RSSM training involves maximizing the Evidence Lower Bound (ELBO) (Hafner et al., 2019b)

$$\log p(\mathbf{o}_{\leq T} \mid \mathbf{a}_{\leq T}) \geq \sum_t \Big(\mathbb{E}_{q(\mathbf{z}_t \mid \mathbf{o}_{\leq t}, \mathbf{a}_{<t})} \left[\log p(\mathbf{o}_t \mid \mathbf{z}_t) + \log p(r_t \mid \mathbf{z}_t) \right] - \mathbb{E}_{q(\mathbf{z}_{t-1} \mid \mathbf{o}_{\leq t-1}, \mathbf{a}_{<t-1})} \left[D_{\text{KL}} \left(q(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1}, \mathbf{o}_t) \,\|\, p(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1}) \right) \right] \Big). \tag{7}$$

DreamerV2/V3 (Hafner et al., 2020; 2023) extend Dreamer (Hafner et al., 2019a) with stability improvements and a categorical RSSM variant (Cat-RSSM), where the stochastic latent state is modeled using discrete categorical variables.

2.3 Uncertainty in Model-Based Reinforcement Learning

Uncertainty is commonly divided into an aleatoric component, describing irreducible environment stochasticity, and an epistemic component, arising from limited data or imperfect learning. In proprioceptive dynamics models, the PE (Lakshminarayanan et al., 2017) represents aleatoric uncertainty as a Gaussian distribution over the next state, while epistemic uncertainty is captured by disagreement among M ensemble members $\{p_i(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)\}_{i=1}^{M}$. Consequently, aligned ensemble predictions indicate a reliable approximation of environment dynamics, assuming sufficient representational capacity of the model (Yu et al., 2020; Frauenknecht et al., 2024).

Similarly, latent dynamics models often use ensembles of latent transition predictors described as $\{g_i(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1})\}_{i=1}^{M}$ to measure disagreement in latent space (Sekar et al., 2020). Alignment of ensemble predictions for a given pair $(\mathbf{z}_{t-1}, \mathbf{a}_{t-1})$ indicates that the latent transition model (Eq. (3)) is well-approximated (Lakshminarayanan et al., 2017), which is typically interpreted as evidence that the latent dynamics reflect true environment dynamics (Sekar et al., 2020; Seo et al., 2025).

3 Related Work

Proprioceptive Dynamics Models. Proprioceptive dynamics models are commonly learned as PE-based predictors over relatively low-dimensional state spaces (Lakshminarayanan et al., 2017). Epistemic uncertainty in these models is used to constrain model exploitation (Yu et al., 2020; Buckman et al., 2018; Frauenknecht et al., 2024; 2025), guide exploration (Sancaktar et al., 2022; Sukhija et al., 2023), or promote caution (Yu et al., 2023; Vignola et al., 2026; Frauenknecht et al., 2026).

Latent Dynamics Models. In vision-based RL, latent dynamics models learn compact representations from high-dimensional image observations. Approaches include reconstruction-based models (Finn & Levine, 2017; Ha & Schmidhuber, 2018; Zhang et al., 2019), self-supervised representation learning methods (Pathak et al., 2017; Schwarzer et al., 2020; Hansen et al., 2022; Zhou et al., 2024), and sequence-based architectures (Chen et al., 2022; Micheli et al., 2022). Among these, the RSSM (Hafner et al., 2019b;a; 2020; 2023) is a widely used and efficient baseline. Uncertainty-aware objectives in latent dynamics models fall into two main categories: explicit epistemic uncertainty estimates used for robustness (Wang et al., 2020; Zhu et al., 2020; Seyde et al., 2020; Filos et al., 2022; Liu et al., 2026), exploration (Sekar et al., 2020; Sancaktar et al., 2025), or safety (Seo et al., 2025; Nakamura et al., 2025); and prediction-based intrinsic signals such as curiosity (Guo et al., 2022; Jarrett et al., 2022). While differing in formulation, both objectives rely on the assumption that uncertainty-related signals in latent space reflect long-term transition reliability. In this work, we revisit this assumption for RSSMs using ensemble-based epistemic uncertainty estimates, which closely align with uncertainty modeling practices in proprioceptive dynamics models.

4 Problem Statement

Epistemic uncertainty quantification via ensemble disagreement is well-established in proprioceptive dynamics models (Yu et al., 2020; Frauenknecht et al., 2024; 2025) and has increasingly been adopted in latent dynamics models, in particular RSSM-based approaches (Seyde et al., 2020; Sekar et al., 2020; Seo et al., 2025), where ensemble disagreement between transition predictors (Sekar et al., 2020) serves as the latent counterpart of proprioceptive PE-based uncertainty. This correspondence makes RSSMs a representative architecture to study whether uncertainty estimation principles transfer from proprioceptive to latent dynamics. In particular, employing ensemble disagreement in latent dynamics implicitly assumes that *local* transition uncertainty remains informative about *global* trajectory-level model error during prolonged rollouts, as observed in proprioceptive dynamics models (Janner et al., 2019). In this work, we empirically revisit this assumption via three questions:

- *(i)* How do RSSM rollouts evolve under in- and out-of-distribution settings compared to proprioceptive dynamics models?
- *(ii)* Does *local* ensemble-based epistemic uncertainty in RSSM latent dynamics remain informative about *global* trajectory-level deviations from true environment dynamics?
- *(iii)* How does the RSSM latent behavior affect reward prediction and therefore policy learning?

Our analysis reveals a failure mode in RSSM latent dynamics where epistemic uncertainty decouples from compounding model error during prolonged rollouts, leading to reward overestimation.

5 Problem Evaluation

We empirically investigate the above questions using RSSM (Hafner et al., 2019b) and Cat-RSSM (Hafner et al., 2020; 2023). Section 5.1 analyzes latent and proprioceptive rollout behavior under a distribution shift, Section 5.2 relates ensemble uncertainty to proprioceptive discrepancy, and Section 5.3 evaluates the impact of latent-environment misalignment on reward prediction.

Experimental Setup. Experiments are conduced on four DMC Suite tasks (Tassa et al., 2018): Cartpole Swingup, Cheetah Run, Hopper Hop, and Walker Run. For each task, we perform $S = 5$

training runs with different random seeds. The RSSM implementation follows Becker & Neumann (2022), while the Cat-RSSM implementation follows Becker et al. (2024), isolating the categorical latent space by omitting other DreamerV2/V3 modifications. All runs are trained for 1 Mio. environment steps. Additional details of our implementation[1] are provided in Appendix 7.

Uncertainty Estimation. We train an external ensemble of $M = 5$ one-step latent transition predictors $\{g_i(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1})\}_{i=1}^{M}$, each optimized via the negative Gaussian log-likelihood of the next latent state predicted by the representation model (Eq. 1). Following Frauenknecht et al. (2024), epistemic uncertainty is computed as the ensemble disagreement over predicted next latent state distributions using Geometric Jensen-Shannon divergence (Nielsen, 2019).

Random Rollouts Collection. For each trained run, we collect three types of rollouts from $N = 1000$ random initial states with rollout length $T = 50$ using the trained greedy policy: (1) *posterior-informed* rollouts (Eq. (6)); (2) *closed-loop prior* rollouts (Eq. (5)), initialized from posterior-informed states; and (3) *open-loop prior* rollouts, using actions from posterior-informed rollouts. Closed-loop refers to rollouts where actions are selected by a reactive policy based on the current latent states, whereas open-loop rollouts follow a predefined action sequence without feedback from the policy. Posterior-informed rollouts represent the most informed rollouts the model can produce, leveraging observation-updated latent states and closed-loop actions while decoupling transition dynamics from observation noise. Closed-loop prior rollouts correspond to latent imagination used during Dreamer's policy learning, whereas open-loop prior rollouts isolate the effects of model stochasticity and error accumulation under a fixed action sequence.

ID and OOD Initial States. To analyze model behavior under distribution shifts, we select in-distribution (ID) and out-of-distribution (OOD) initial states from the post-training rollouts under the learned policy. We use ensemble disagreement over latent states and actions as a measure of *local* uncertainty, which has been shown to reliably capture epistemic transition uncertainty under distribution shifts (Lakshminarayanan et al., 2017; Chua et al., 2018). ID states are selected as the K lowest-uncertainty states from the posterior-informed and closed-loop prior rollouts. OOD states are selected as the K highest-uncertainty states from the posterior-informed and open-loop prior rollouts, where fixed actions allow model stochasticity and error accumulation to progressively induce a state-action mismatch. We use $K = 10$ initial states per seed, task, and setup (ID/OOD).

5.1 Attractor Evaluation

To address *Question (i)*, we analyze how ID and OOD initial states affect rollout behavior in proprioceptive and latent dynamics models. Under a distribution shift, proprioceptive dynamics models typically exhibit prediction instabilities due to compounding model error (Janner et al., 2019). We investigate whether RSSMs exhibit a similar behavior.

Visualization of Dynamical Behavior. To avoid attributing model behavior to suboptimal training, we select the best-performing run for each task, which has the highest mean evaluation return over the training. We embed its posterior-informed and closed-loop prior rollouts into a 2D PCA space of normalized deterministic features $\mathbf{f}_t := (\mathbf{h}_t, \mathbf{z}_t^{\mathrm{m}})$, where $\mathbf{z}_t^{\mathrm{m}}$ denotes the mean stochastic state in RSSM and the categorical mode in Cat-RSSM. We construct a vector field by binning one-step displacements $\hat{\mathbf{f}}_t - \hat{\mathbf{f}}_{t-1}$, where $\mathrm{PCA}(\mathbf{f}_t) = \hat{\mathbf{f}}_t$, and overlay closed-loop rollouts initialized from ID states with lowest and OOD states with the highest mean uncertainty over the rollout horizon.

Proprioceptive Dynamics. We additionally analyze PE dynamics from Infoprop (Frauenknecht et al., 2025) in the MuJoCo HalfCheetah environment (Todorov et al., 2012), applying the same PCA procedure to proprioceptive states. ID and OOD rollouts are closed-loop rollouts with the lowest and highest mean uncertainty over the rollout horizon, respectively, selected from random initial states.

[1] https://github.com/jberger999/biased-dreams

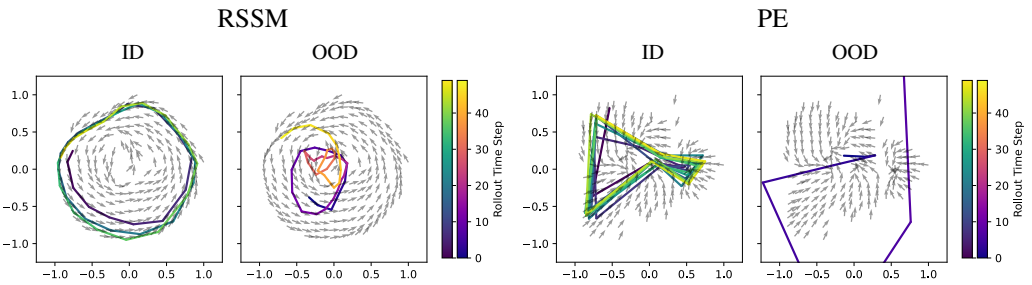

Figure 2: Attractor evaluation for RSSM and PE under ID and OOD initial states on Cheetah Run and Halfcheetah, respectively. Vector fields show one-step displacements in the 2D PCA spaces of latent and proprioceptive states. PE exhibits unstable OOD predictions, whereas RSSM is drawn toward dominant transition flows, indicating an attractor bias toward well-supported latent regions.

As visualized in Fig. 2, the PE exhibits unstable OOD predictions, reflecting compounding model error. In contrast, the RSSM OOD rollout converges toward dominant transition flows, indicating an attractor effect that biases rollouts toward well-supported latent regions. This behavior suggests that distribution shifts manifest differently in RSSMs and proprioceptive dynamics, as latent representations may constrain rollouts toward training-supported regions through transitions that are not necessarily physically accurate. A similar behavior can be observed across other DMC Suite tasks and for Cat-RSSM, as reported in Appendix 8.1. However, given the qualitative nature of this analysis, these findings should be interpreted as observations rather than definitive findings.

5.2 Proprioceptive Discrepancy

We evaluate proprioceptive discrepancy between RSSM latent rollouts and true environment behavior under ID/OOD initial states, and relate it to ensemble-based epistemic uncertainty, addressing *Question (ii)*. While ensemble disagreement reliably captures local epistemic uncertainty, we investigate whether this extends to *global* trajectory-level prediction accuracy.

Proprioceptive State Decoder. To compare latent predictions and true proprioceptive dynamics, we augment the RSSM with a proprioceptive state decoder (PD) $p(\mathbf{s}_t \mid \mathbf{z}_t)$, trained jointly via the ELBO (Eq. (7)) to predict transformed proprioceptive states excluding the agent's x-position. Although the decoder shapes latent representations during training, proprioceptive reconstruction quality remains limited without an explicit structural objective (Peper et al., 2025; Gupta et al., 2021). As shown in Fig. 3, adding the PD has negligible impact on Dreamer's performance.

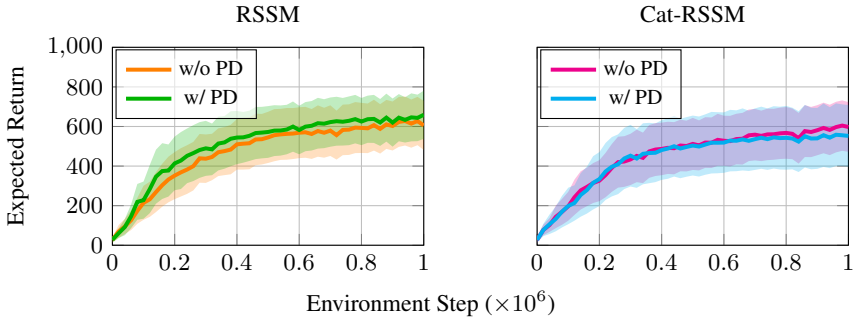

Figure 3: Dreamer's performance with and without a proprioceptive state decoder (PD) for RSSM and Cat-RSSM across four DMC Suite tasks. The addition of a PD has negligible impact on policy learning, with both architectures showing comparable performance to their respective no-PD variants. Results are averaged over 5 seeds per setting; shaded areas indicate 95% confidence intervals.

Latent Dynamics Rollouts. For each of the K ID and OOD initial states, we evaluate R closed-loop posterior-informed and R closed-loop prior rollouts of length $T = 50$. These rollout types provide complementary views of the learned dynamics: Closed-loop posterior-informed rollouts leverage observation-updated latent states for the most informed transition prediction, whereas closed-loop prior rollouts evaluate predictive behavior during imagination. Actions are selected greedily, while dynamics predictions remain stochastic. We set the number of rollouts to $R = 10$.

Computation of Proprioceptive Discrepancy. Proprioceptive discrepancy is measured as the mean ℓ_1 distance between predicted and ground-truth body positions, where ground truth is obtained by simulating the predicted action sequence from a given proprioceptive start state.

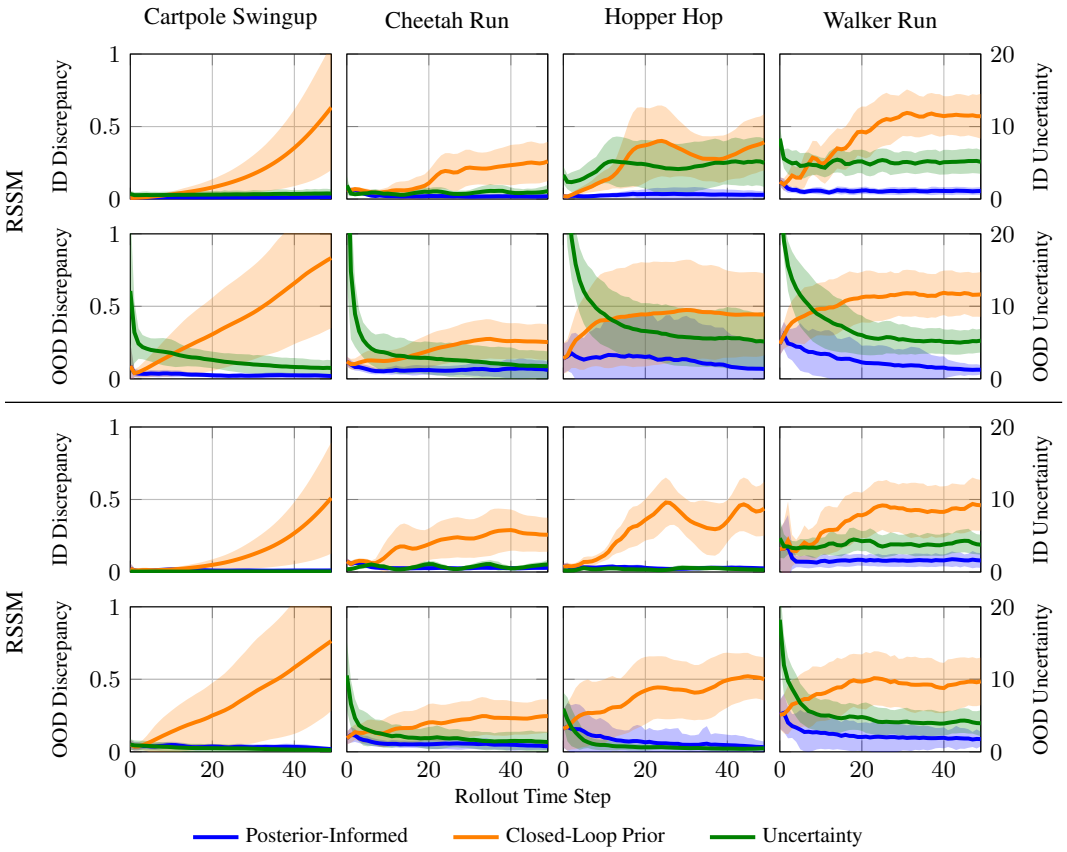

Figure 4: Proprioceptive discrepancy versus prior uncertainty for ID and OOD initial states in RSSM and Cat-RSSM. Solid lines show means over $R \cdot K \cdot S = 500$ samples per time step and setting; shaded areas denote standard deviation. Although uncertainty captures initial distribution shifts, it does not track the compounding proprioceptive errors that emerge during latent rollouts.

We report the stepwise mean and standard deviation of proprioceptive discrepancy and prior uncertainty (Fig. 4) and discuss RSSM and Cat-RSSM jointly due to their similar behavior. Posterior-informed rollouts remain accurate for ID initial states, while initially high OOD discrepancy decreases as observations refine the latent state. Closed-loop prior rollouts accumulate proprioceptive discrepancy in both settings, but uncertainty does not reflect this error growth: ID uncertainty remains low and stable, whereas OOD uncertainty starts elevated, but rapidly converges to ID levels. This behavior can be attributed to the attractor effect observed for OOD rollouts in Section 5.1. While ensemble-based uncertainty appears locally informative, RSSM latent dynamics favor regions that are well-supported by the training distribution and exhibit low ensemble disagreement during prolonged rollouts, preventing uncertainty from reliably tracking compounding model error.

5.3 Reward Evaluation

To address *Question (iii)*, we evaluate how latent-environment misalignment affects reward prediction. We compute a stepwise signed reward discrepancy $r_t^{\text{pred}} - r_t^{\text{sim}}$ over the collected closed-loop prior and posterior-informed rollouts, where r_t^{pred} denotes the model-predicted reward along the latent rollout and r_t^{sim} the reward obtained by executing the same action sequence in the environment.

We report the stepwise mean and standard deviation of signed reward discrepancy (Fig. 5) and discuss RSSM and Cat-RSSM jointly due to their similar behavior. Posterior-informed rollouts accurately predict rewards under observation updates, whereas closed-loop prior rollouts systematically overestimate them. Together with the attractor behavior observed in Section 5.1, this suggests that latent rollouts are biased toward well-supported regions that appear to coincide with high-performing behaviors in dense-reward tasks, contributing to overly optimistic reward predictions.

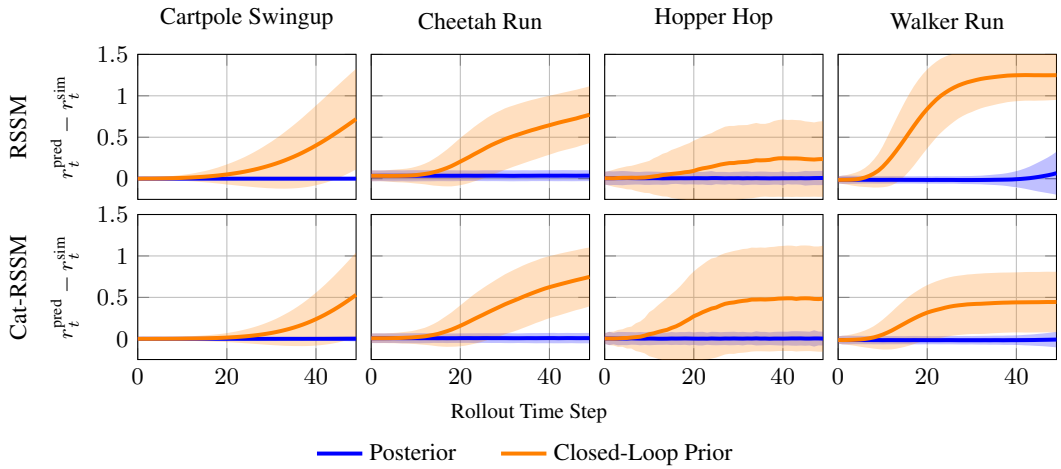

Figure 5: Signed reward discrepancy in RSSM and Cat-RSSM. Solid lines show means over $N \cdot S = 5000$ samples per time step and setting; shaded areas denote standard deviation. Closed-loop prior rollouts systematically overestimate rewards, consistent with an attractor bias toward well-supported latent regions that appear to coincide with high-performing behaviors.

6 Conclusion

Our findings reveal a structural limitation of epistemic uncertainty estimation in RSSMs: While ensemble disagreement has been shown to capture *local* transition reliability (Lakshminarayanan et al., 2017; Chua et al., 2018), this relationship does not necessarily extend to *global* trajectory-level model error during prolonged latent rollouts. We provide evidence for an attractor behavior in RSSM latent dynamics, where rollouts are biased toward latent regions well-supported by the learned dynamics, as indicated by low ensemble disagreement. As a result, epistemic uncertainty can diminish despite increasing discrepancies from true environment dynamics, masking compounding model error. In dense-reward tasks, this behavior may contribute to overly optimistic reward predictions, as attractor regions appear to align with high-reward behaviors.

Future Work. We hypothesize that this limitation originates from inductive biases in VAE-based latent dynamics and cannot be resolved through improved uncertainty estimation alone. Instead, addressing it may require changes to latent dynamics modeling, e.g., through latent space restructuring (Li et al., 2024; Becker et al., 2024), more principled variational inference (Becker & Neumann, 2022), or alternative latent parametrizations. More broadly, our findings motivate further investigation of epistemic uncertainty in latent dynamics models beyond RSSMs. A formal characterization of the attractor mechanism remains an important direction for future work.

Acknowledgments

We thank Tim Elsner and Leif Kobbelt for their support during an earlier research project that inspired this paper. This work was partially supported by the Robotics Institute Germany (RIG). The authors gratefully acknowledge the computing time provided to them at the NHR Center NHR4CES at RWTH Aachen University (project number p0022301). This is funded by the Federal Ministry of Research, Technology and Space, and the state governments participating on the basis of the resolutions of the GWK for national high performance computing at universities (www.nhr-verein.de/unsere-partner).

References

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Andrew Barto and Richard Sutton. *Reinforcement learning: an introduction*. The MIT Press, 2018.

Philipp Becker and Gerhard Neumann. On uncertainty in deep state space models for model-based reinforcement learning. *Transactions on Machine Learning Research (TMLR)*, 2022.

Philipp Becker, Sebastian Mossburger, Fabian Otto, and Gerhard Neumann. Combining reconstruction and contrastive methods for multimodal representations in rl. *Reinforcement Learning Conference (RLC)*, 2024.

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

Jacob Buckman, Danijar Hafner, George Tucker, Eugene Brevdo, and Honglak Lee. Sample-efficient reinforcement learning with stochastic ensemble value expansion. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018.

Chang Chen, Yi-Fu Wu, Jaesik Yoon, and Sungjin Ahn. Transdreamer: Reinforcement learning with transformer world models. *arXiv preprint arXiv:2202.09481*, 2022.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018.

Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de Las Casas, et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897):414–419, 2022.

Marc Deisenroth and Carl E Rasmussen. Pilco: A model-based and data-efficient approach to policy search. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 465–472, 2011.

Angelos Filos, Eszter Vértes, Zita Marinho, Gregory Farquhar, Diana Borsa, Abram Friesen, Feryal Behbahani, Tom Schaul, André Barreto, and Simon Osindero. Model-value inconsistency as a signal for epistemic uncertainty. *International Conference on Learning Representations (ICLR)*, 2022.

Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *2017 IEEE international conference on robotics and automation (ICRA)*, pp. 2786–2793. IEEE, 2017.

Bernd Frauenknecht, Artur Eisele, Devdutt Subhasish, Friedrich Solowjow, and Sebastian Trimpe. Trust the model where it trusts itself–model-based actor-critic with uncertainty-aware rollout adaption. *International Conference on Machine Learning (ICML)*, 2024.

Bernd Frauenknecht, Devdutt Subhasish, Friedrich Solowjow, and Sebastian Trimpe. On rollouts in model-based reinforcement learning. *International Conference on Learning Representations (ICLR)*, 2025.

Bernd Frauenknecht, Lukas Kesper, Daniel Mayfrank, Henrik Hose, and Sebastian Trimpe. Uncertainty-aware predictive safety filters for probabilistic neural network dynamics. *Reinforcement Learning Conference*, 2026.

Zhaohan Guo, Shantanu Thakoor, Miruna Pîslar, Bernardo Avila Pires, Florent Altché, Corentin Tallec, Alaa Saade, Daniele Calandriello, Jean-Bastien Grill, Yunhao Tang, et al. Byol-explore: Exploration by bootstrapped prediction. *Advances in neural information processing systems*, 35: 31855–31870, 2022.

Rushil Gupta, Vishal Sharma, Yash Jain, Yitao Liang, Guy Van den Broeck, and Parag Singla. Towards an interpretable latent space in structured models for video prediction. *arXiv preprint arXiv:2107.07713*, 2021.

David Ha and Jürgen Schmidhuber. World models. *Conference on Neural Information Processing Systems (NeurIPS)*, 2(3):440, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019a.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning (ICML)*, pp. 2555–2565. PMLR, 2019b.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020.

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

Nicklas Hansen, Xiaolong Wang, and Hao Su. Temporal difference learning for model predictive control. *International Conference on Machine Learning (ICML)*, 2022.

Nicklas Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. *International Conference on Learning Representations (ICLR)*, 2024.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

Daniel Jarrett, Corentin Tallec, Florent Altché, Thomas Mesnard, Rémi Munos, and Michal Valko. Curiosity in hindsight: Intrinsic exploration in stochastic environments. *arXiv preprint arXiv:2211.10515*, 2022.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

Sergey Levine and Vladlen Koltun. Guided policy search. In *International Conference on Machine Learning (ICML)*, pp. 1–9. PMLR, 2013.

Chenhao Li, Elijah Stanger-Jones, Steve Heim, and Sangbae Kim. Fld: Fourier latent dynamics for structured motion representation and learning. *International Conference on Learning Representations (ICLR)*, 2024.

Zhenxian Liu, Peixi Peng, Yangru Huang, and Yonghong Tian. Perceiving the knowledge boundary: Uncertainty-guided exploration and imagination for world models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 2026.

Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. *arXiv preprint arXiv:2209.00588*, 2022.

Anusha Nagabandi, Gregory Kahn, Ronald S Fearing, and Sergey Levine. Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7559–7566. IEEE, 2018.

Kensuke Nakamura, Lasse Peters, and Andrea Bajcsy. Generalizing safety beyond collision-avoidance via latent-space reachability analysis. *arXiv preprint arXiv:2502.00935*, 2025.

Frank Nielsen. On the jensen–shannon symmetrization of distances relying on abstract means. *Entropy*, 21(5):485, 2019.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pp. 2778–2787. PMLR, 2017.

Jordan Peper, Zhenjiang Mao, Yuang Geng, Siyuan Pan, and Ivan Ruchkin. Four principles for physically interpretable world models. *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

Cansu Sancaktar, Sebastian Blaes, and Georg Martius. Curious exploration via structured world models yields zero-shot object manipulation. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:24170–24183, 2022.

Cansu Sancaktar, Christian Gumbsch, Andrii Zadaianchuk, Pavel Kolev, and Georg Martius. Sensei: Semantic exploration guided by foundation models to learn versatile world models. *arXiv preprint arXiv:2503.01584*, 2025.

Max Schwarzer, Ankesh Anand, Rishab Goel, R Devon Hjelm, Aaron Courville, and Philip Bachman. Data-efficient reinforcement learning with self-predictive representations. *arXiv preprint arXiv:2007.05929*, 2020.

Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to explore via self-supervised world models. In *International Conference on Machine Learning (ICML)*, pp. 8583–8592. PMLR, 2020.

Junwon Seo, Kensuke Nakamura, and Andrea Bajcsy. Uncertainty-aware latent safety filters for avoiding out-of-distribution failures. *Conference on Robot Learning (CoRL)*, 2025.

Tim Seyde, Wilko Schwarting, Sertac Karaman, and Daniela Rus. Learning to plan optimistically: Uncertainty-guided deep exploration via latent model ensembles. *Computing Research Repository (CoRR)*, 2020.

Bhavya Sukhija, Lenart Treven, Cansu Sancaktar, Sebastian Blaes, Stelian Coros, and Andreas Krause. Optimistic active exploration of dynamical systems. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 38122–38153. Curran Associates, Inc., 2023.

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

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012.

Luca Vignola, Bruce D Lee, Manish Prajapat, Manuel Wendl, Melanie Zeilinger, Andreas Krause, and Yarden As. Sampling-based safe reinforcement learning. *ICML Workshop on Decision-Making from Offline Datasets to Online Adaptation*, 2026.

Junjie Wang, Qichao Zhang, Dongbin Zhao, Mengchen Zhao, and Jianye Hao. Dynamic horizon value estimation for model-based reinforcement learning. *arXiv preprint arXiv:2009.09593*, 2020.

Grady Williams, Nolan Wagener, Brian Goldfain, Paul Drews, James M Rehg, Byron Boots, and Evangelos A Theodorou. Information theoretic mpc for model-based reinforcement learning. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1714–1721. IEEE, 2017.

Dongjie Yu, Wenjun Zou, Yujie Yang, Haitong Ma, Shengbo Eben Li, Yuming Yin, Jianyu Chen, and Jingliang Duan. Safe model-based reinforcement learning with an uncertainty-aware reachability certificate. *IEEE Transactions on Automation Science and Engineering*, 21(3):4129–4142, 2023.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:14129–14142, 2020.

Marvin Zhang, Sharad Vikram, Laura Smith, Pieter Abbeel, Matthew Johnson, and Sergey Levine. Solar: Deep structured representations for model-based reinforcement learning. In *International conference on machine learning*, pp. 7444–7453. PMLR, 2019.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *International Conference on Machine Learning (ICML)*, 2024.

Guangxiang Zhu, Minghao Zhang, Honglak Lee, and Chongjie Zhang. Bridging imagination and reality for model-based deep reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:8993–9006, 2020.

Supplementary Materials

7 Experiment Setup

Our evaluation code[2] follows Becker & Neumann (2022); Becker et al. (2024), which in turn builds on Hafner et al. (2019a; 2020). For details on Infoprop, see Frauenknecht et al. (2025).

7.1 Training Details

Environments. All environments use an action repeat of 2. Images are 64×64 pixels with 5-bit color depth, following Hafner et al. (2019b).

Architecture. For RSSM, we use a stochastic state size of 30 and a deterministic state size for 200. For Cat-RSSM, the stochastic state is replaced by 32 categoricals with 32 classes each. Independent of architecture, all layers have 300 units, with standard encoders and decoders using ReLU. Note that Cat-RSSM, as proposed by Hafner et al. (2020), also increases model size, which we omit to ensure a fair comparison with RSSM. The transition matrix output is transformed via a sigmoid-based function, as described in Becker & Neumann (2022).

Proprioceptive Decoder and Ensemble. The proprioceptive state decoder follows the decoder architecture in Becker & Neumann (2022), with 3 layers of size 300 and ELU activation. Each angle θ in the proprioceptive state is encoded as $(\sin\theta, \cos\theta)$ and reconstructed for evaluation and simulation from $\arctan2(y_1, y_2)$ given network angle predictions (y_1, y_2). Each member of the ensemble has 5 linear layers of size 300, intermediately normalized with LayerNorm, and activated with ELU. The ensemble predicts the next deterministic state, following standard ensemble-based implementations (Seo et al., 2025).

Loss. RSSM and actor-critic losses follow Hafner et al. (2019a), including a separate reconstruction loss for the proprioceptive decoder, when applicable. The ensemble loss is the negative Gaussian log-likelihood of the target state. During ensemble loss computation, inputs and targets are detached to prevent gradient flow through the model and actor. For Cat-RSSM, we additionally employ KL balancing as proposed by Hafner et al. (2020), separately scaling posterior and prior entropy in the KL term with $\alpha = 0.8$. We omit adding action entropy in the policy loss to focus on architectural differences between RSSM and Cat-RSSM, rather than differences in the learning objective.

Training. We begin training with 5 random episodes and collect one additional episode every 100 model update steps using exploration noise of 0.3. Each update samples 50 subsequences of length 50 uniformly from all collected data. Adam is used with learning rates and gradient clipping as in Hafner et al. (2019a): latent dynamics, $6 \cdot 10^{-4}$; actor-critic, $8 \cdot 10^{-5}$; gradient norm clipped at 100. The ensemble uses a separate optimizer with the same optimizer hyperparameters as the latent dynamics model.

7.2 Evaluation Details

Aggregation. For each DMC Suite task, we perform $S = 5$ training runs with different random seeds to account for training variability while keeping computational cost manageable. For the proprioceptive and reward discrepancy plots (Figs. 4 and 5), we report the stepwise mean and standard deviation over all collected samples. As generic descriptive statistics, mean and standard deviation summarize the overall tendency and variability of rollout distributions without imposing a specific distributional model.

[2] https://github.com/jberger999/biased-dreams

Rollout Types. Posterior-informed rollouts are obtained by applying a one-step prior transition (Eq. 6) from each posterior state inferred by the representation model (Eq. 1). Compared to posterior rollouts, this reduces observation-induced noise in proprioceptive reconstructions, enabling a cleaner assessment of predictive dynamics accuracy. Closed-loop prior rollouts start from a posterior state and evolve under a reactive greedy policy with stochastic latent dynamics. Open-loop prior rollouts instead replay the action sequence of a posterior rollout under stochastic latent dynamics, isolating the effects of model stochasticity and error accumulation. In practice, these rollouts tend to induce state-action mismatches, making them well-suited for identifying highly uncertain transitions.

8 Additional Results

8.1 Attractor Evaluation

The remaining DMC Suite environments exhibit similar latent rollout patterns in the embedded PCA space consistent with those observed in the Cheetah Run evaluation presented in Section 5.1, across both RSSM and Cat-RSSM (Fig. 6). These observations indicate that the observed attractor behavior is not specific to a single task or architecture, but appears consistently across all evaluated settings.

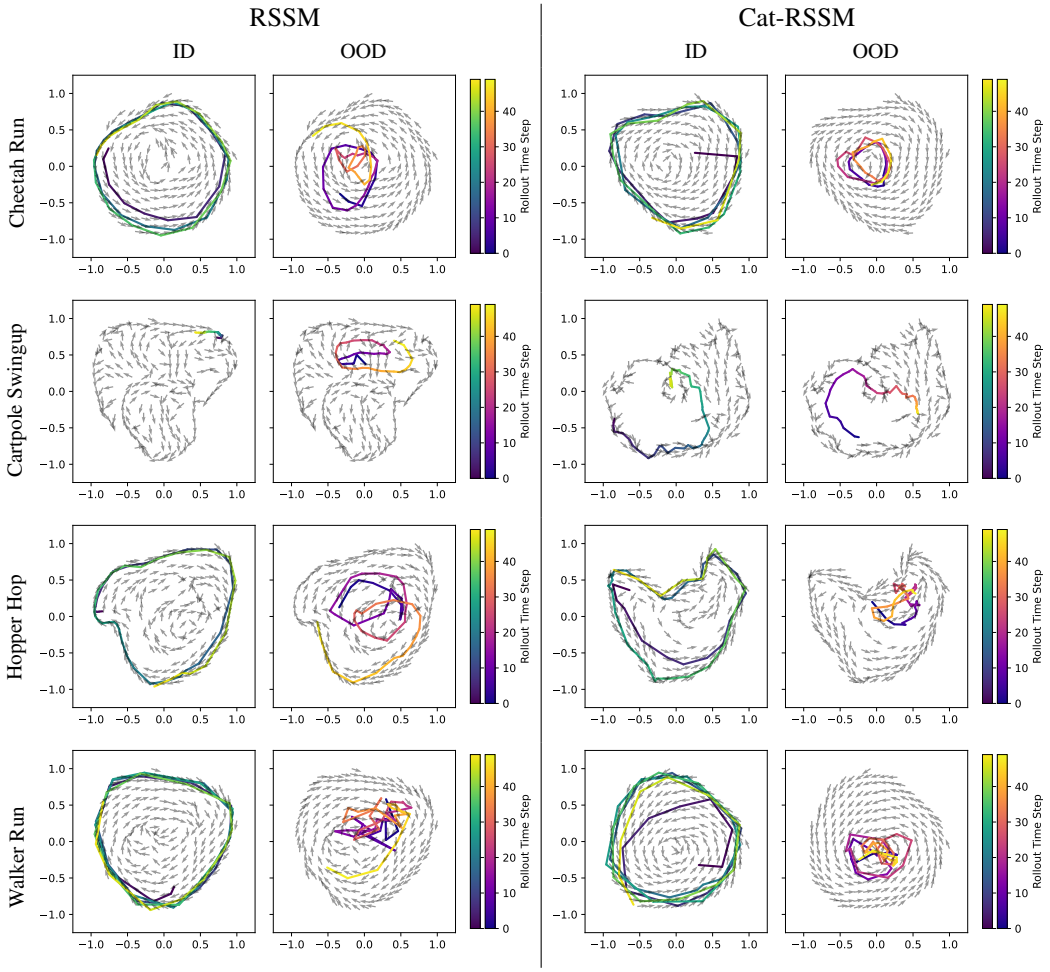

Figure 6: Attractor analysis of RSSM and Cat-RSSM across DMC Suite environments.

8.2 Proprioceptive Discrepancy

We report Dreamer's RL performance using RSSM and Cat-RSSM, with and without a PD, separately for each task (Fig. 7). The per-task results are consistent with the aggregated plots in Section 5.2, showing that the incorporating the PD into the latent dynamics model has negligible impact on Dreamer's performance.

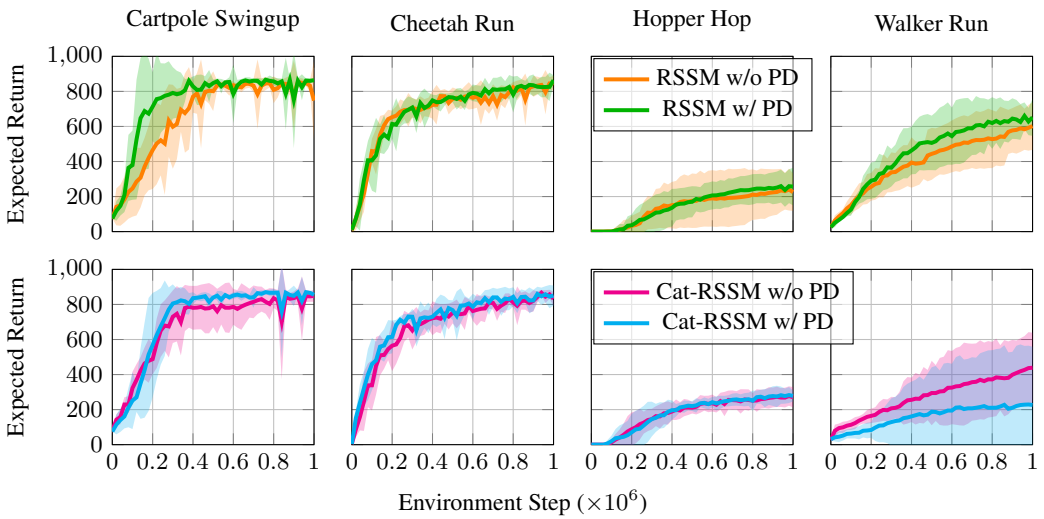

Figure 7: Dreamer's performance is largely unchanged by the use of the proprioceptive state decoder (PD) between approaches. Results are averaged over 5 seeds; shaded areas show 95% confidence intervals.

8.3 Image and Proprioceptive Reconstructions

Fig. 8 visualizes the proprioceptive state rollouts predicted by Infoprop's PE, when simulated in the environment, corresponding to the exemplary ID and OOD cases discussed in the main text.

Figs. 9, 10 and Figs. 11, 12 compare simulator ground truth with exemplary image and proprioceptive rollouts, respectively, reconstructed from the corresponding posterior-informed and closed-loop prior rollouts. We visualize the first posterior-informed and closed-loop prior rollouts from our collection. Overall, the reconstructed image and proprioceptive state rollouts are largely consistent with each other, indicating that the proprioceptive decoder achieves reconstruction quality comparable to the image decoder. Posterior-informed rollouts remain close to the ground truth, whereas closed-loop prior rollouts exhibit increasing misalignment between predicted states and environment dynamics over longer horizons, independent of initial state setting (ID/OOD).

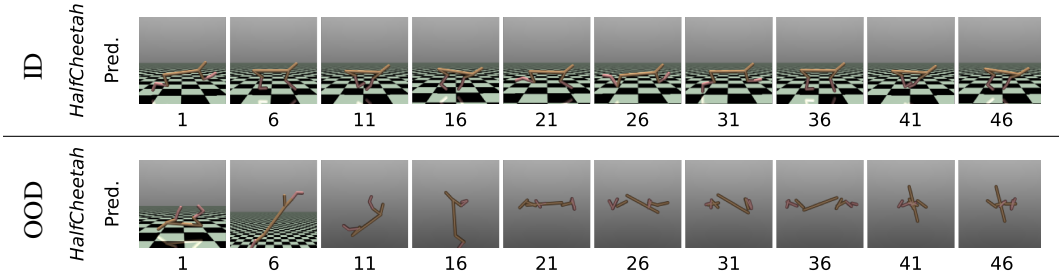

Figure 8: Comparison of predicted rollouts in ID and OOD settings for Infoprop's PE.

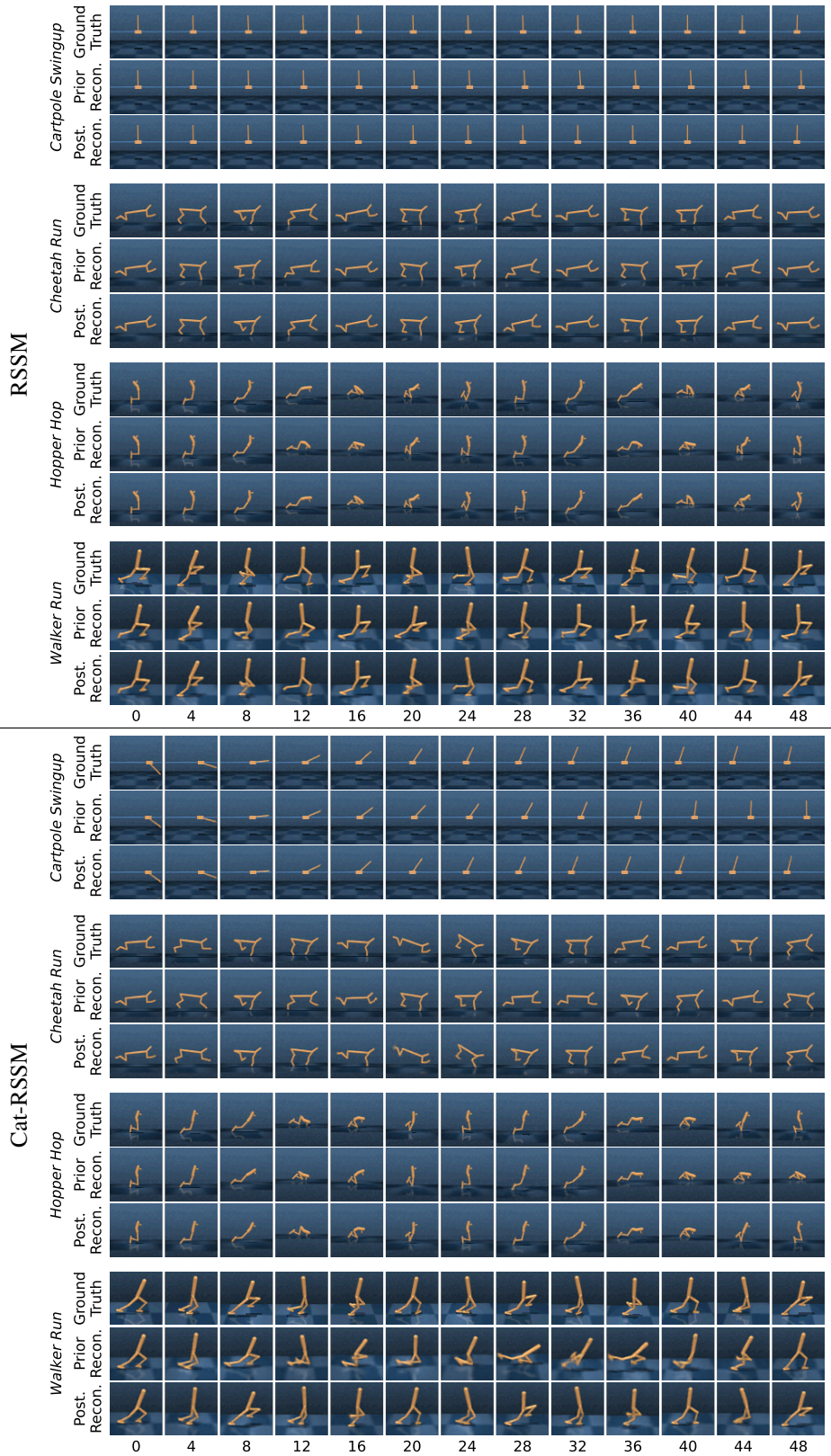

Figure 9: Step-wise comparison of reconstructed posterior-informed and closed-loop prior image rollouts in ID setting for RSSM and Cat-RSSM.

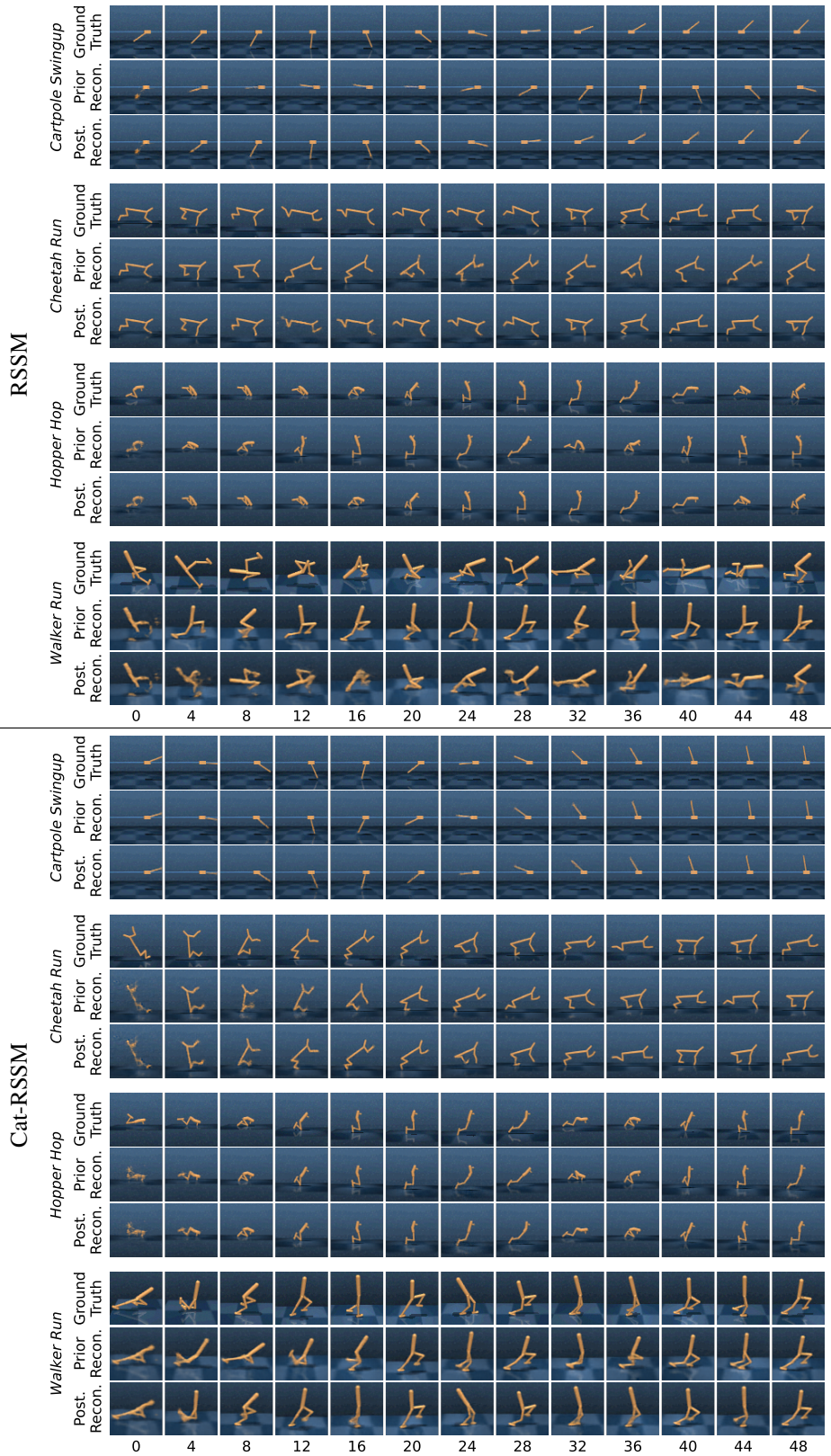

Figure 10: Step-wise comparison of reconstructed posterior-informed and closed-loop prior image rollouts in OOD setting for RSSM and Cat-RSSM.

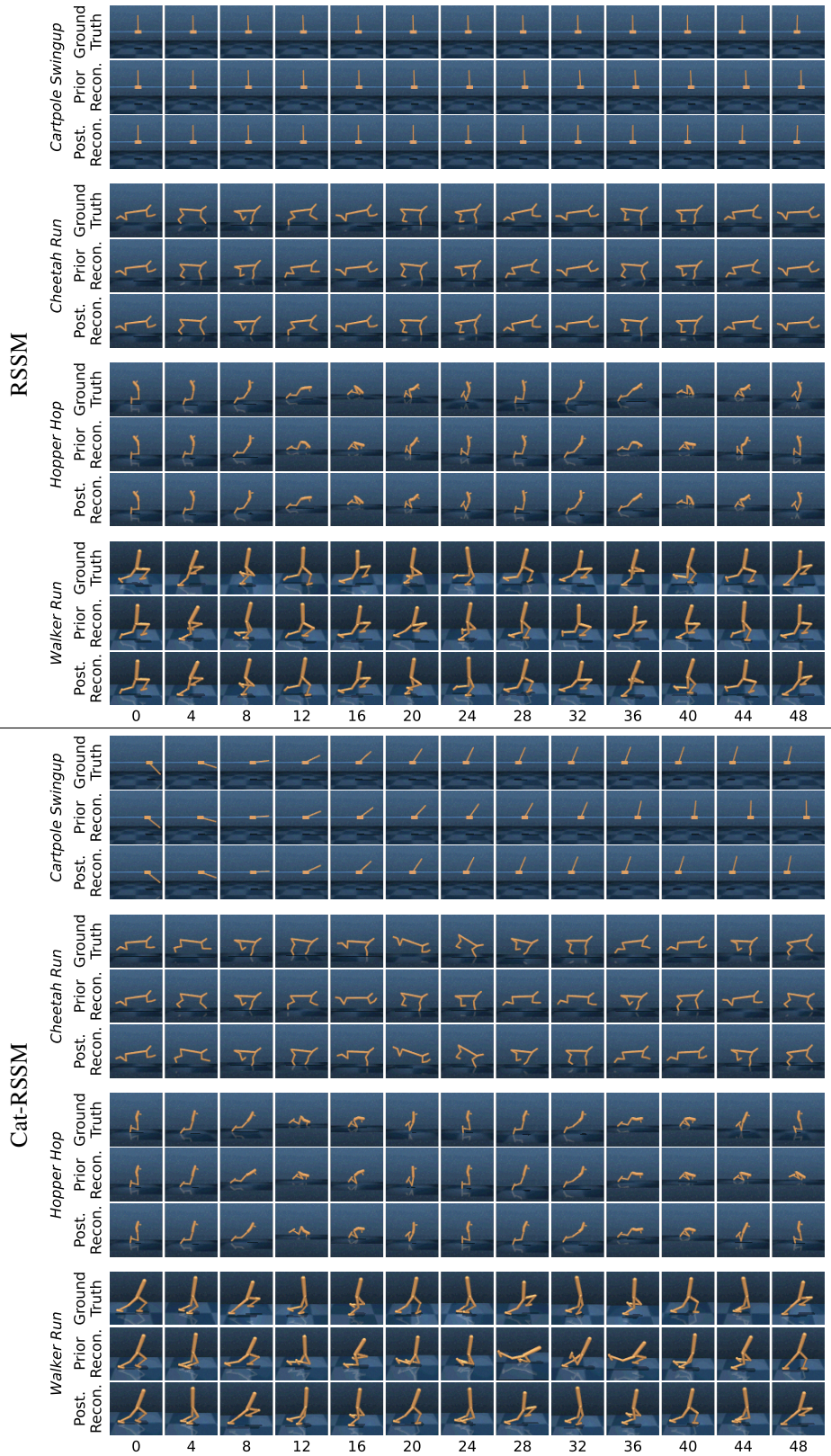

Figure 11: Step-wise comparison of reconstructed posterior-informed and closed-loop prior proprioceptive rollouts in ID setting for RSSM and Cat-RSSM.

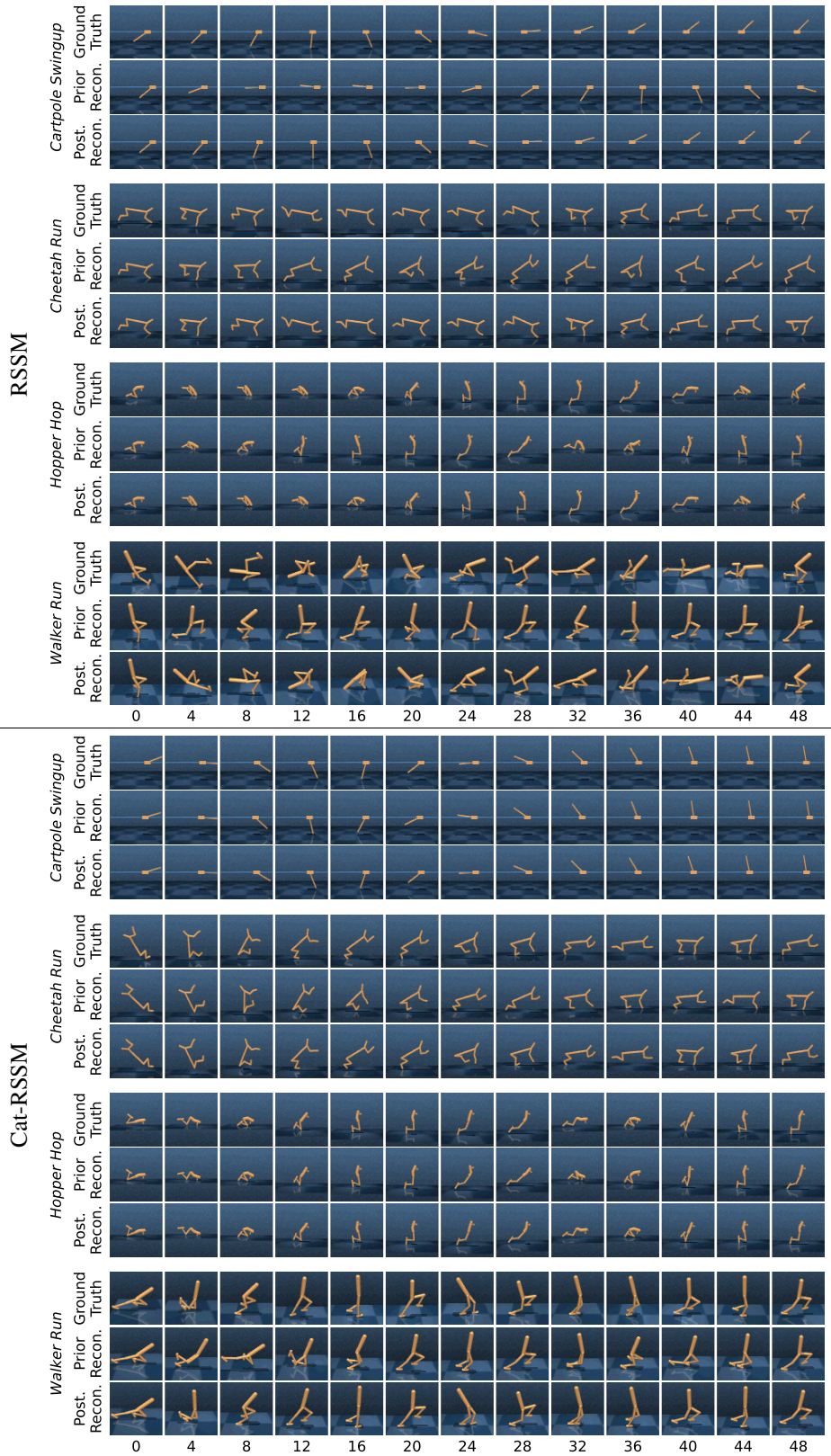

Figure 12: Step-wise comparison of reconstructed posterior-informed and closed-loop prior proprioceptive rollouts in OOD setting for RSSM and Cat-RSSM.