UNISafe: Uncertainty-aware Latent Safety Filters for Avoiding Out-of-Distribution Failures

Junwon Seo, Kensuke Nakamura, Andrea Bajcsy
Carnegie Mellon University
{junwonse, kensuken, abajcsy}@andrew.cmu.edu

***Abstract*—Recent advances in generative world models have enabled classical safe control methods, such as Hamilton-Jacobi (HJ) reachability, to generalize to complex robotic systems operating directly from high-dimensional sensor observations. However, obtaining comprehensive coverage of all safety-critical scenarios during world model training is extremely challenging. As a result, latent safety filters built on top of these models may miss novel hazards and even fail to prevent known ones, overconfidently misclassifying risky out-of-distribution (OOD) situations as safe. To address this, we introduce an uncertainty-aware latent safety filter that proactively steers robots away from both known and unseen failures. Our key idea is to use the world model's epistemic uncertainty as a proxy for identifying unseen potential hazards. We propose a principled method to detect OOD world model predictions by calibrating an uncertainty threshold via conformal prediction. By performing reachability analysis in an augmented state space–spanning both the latent representation and the epistemic uncertainty–we synthesize a latent safety filter that can reliably safeguard arbitrary policies from both known and unseen safety hazards. In simulation and hardware experiments on vision-based control tasks with a Franka manipulator, we show that our uncertainty-aware safety filter preemptively detects potential unsafe scenarios and reliably proposes safe, in-distribution actions. Video results can be found on the project website: https://cmu-intentlab.github.io/UNISafe**

I. Introduction

Robots operating in complex open-world environments must interact safely with the world based on high-dimensional sensor observations. A promising approach to scale safe control to such settings is to learn a world model (WM) [19] that jointly compresses observations into compact latent representations and predicts their dynamics, allowing the robot to anticipate the consequences of candidate actions to prevent unsafe ones [41]. However, without unlimited unsafe exploration, the WM's training data can fail to capture the full range of possible safety hazards. For example, in the *Jenga* game (right, Fig. 1), most of the ways in which the tower can fall are not seen during training. During interaction, if the robot fails to reliably predict how its actions can lead to such *out-of-distribution (OOD)* scenarios, it may inadvertently execute actions that lead to unsafe outcomes [2, 68].

One way to address this model uncertainty is through OOD detection, which identifies when the robot encounters anomalous observations or generates uncertain predictions [55, 50, 69]. However, on its own, OOD detection lacks actionable mitigation strategies, leaving robots *aware* of their uncertainty yet unable to *act* appropriately. Here, safe control methods such as Hamilton-Jacobi (HJ) reachability analysis [39, 63] offer a complementary approach by synthesizing fallback policies that proactively enforce safety constraints, keeping the system within control-invariant sets. Yet, they typically assume a perfect state representation and a faithful dynamics model, assumptions that may not hold in OOD scenarios when relying on a world model for safe control. To bridge this gap, we argue that safety constraints for latent-space control should be augmented to identify unreliable model predictions, enabling the synthesis of a *safety filter* that prevents the system from entering both *known failures* and potentially unsafe *OOD failures*.

In this work, we propose ***UNISafe*** (***UN****certainty-aware* ***I****magination for* ***Safe****ty filtering*): a policy-agnostic safety mechanism that reliably steers robots away from known and unseen safety hazards using a latent world model [19, 20]. Our key idea is to use the world model's epistemic uncertainty as a proxy for identifying unseen potential hazards. We propose a principled method to quantify the epistemic uncertainty of the world model and detect unreliable world model predictions by calibrating an uncertainty threshold via conformal prediction. By performing reachability analysis in an augmented state space spanning both the latent states and the uncertainty, we synthesize a safety filter that can reliably prevent a system from entering both predictable and unforeseen failure modes.

We evaluate our framework in simulation and hardware on three vision-based safe-control tasks. We find that ***UNISafe*** effectively prevents failures with world models trained on an offline dataset with limited coverage. Importantly, by penalizing overly optimistic safety evaluations of OOD scenarios, our safety filter preemptively detects potential safety risks and proposes reliable backup actions, consistently guiding the system toward safe, in-distribution behaviors.

II. Related Works

Out-of-distribution Detection for Robotics. Data-driven control often exhibits unreliable behavior when encountering data that deviates from its training distribution [50, 55, 2, 56, 68]. To detect such out-of-distribution (OOD) conditions, uncertainty is estimated via pre-trained feature spaces [67, 35], reconstruction [46, 65, 49, 51], density estimation [13, 36, 28, 8], or ensembles [29, 71, 43, 53]. While these methods can detect OOD and serve as runtime monitors [10, 52, 38, 2], they often lack control invariance, limiting them to passive detection rather than proactive failure prevention. Moreover, they typically do not distinguish between epistemic uncertainty

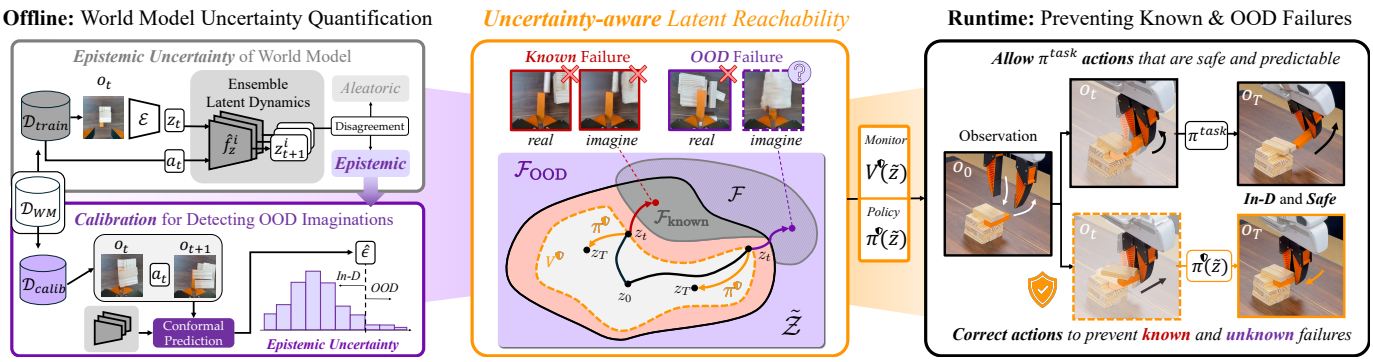

Fig. 1: *Left*: We quantify the world model's epistemic uncertainty in latent space and calibrate an uncertainty threshold via conformal prediction, resulting in an *OOD failure set*, $\mathcal{F}_{\text{OOD}}$. *Center*: Uncertainty-aware latent reachability analysis synthesizes a safety monitor V^{π} and fallback policy π^{π} that steers the system away from both known and OOD failures. *Right*: Our safety filter reliably safeguards arbitrary task policies during hard-to-model vision-based tasks, like a teleoperator playing Jenga.

(i.e., lack of knowledge) and aleatoric uncertainty (i.e., inherent noise) [29, 71, 43], while capturing epistemic uncertainty is critical for reliable OOD detection [54, 30]. To bridge this gap, we quantify the epistemic uncertainty of a world model [54, 61, 30] to formulate a constraint, enabling reachability analysis to synthesize control strategies that prevent the system from OOD scenarios.

Safety Filtering. Safety filtering is a control-theoretic approach for safeguarding robotic systems from unsafe conditions [39, 3, 14, 63, 24, 17]. While they can provide robust safety assurances under model uncertainty [14, 22, 25, 64, 11], they focus on worst-case disturbances, addressing aleatoric uncertainty rather than epistemic uncertainty of the model. Self-supervised [6] and reinforcement learning methods [15, 23] have been used to scale safety filtering to high-dimensional systems, but these approaches typically rely on known system dynamics with simple safety specifications [25, 9] or online rollouts in simulators [26, 21, 42]. To generalize safety filters with complex dynamics and constraints, latent world models [19] have been used [8, 66, 41], but the epistemic uncertainty of the learned model can compromise reliability [43]. Recent works prevent the system from entering OOD states [28, 8], but they restrict in-distribution to safe trajectories or do not construct constraints with calibrated OOD detection [48, 8, 34], limiting their scalability to complex settings. Our method leverages calibrated OOD detection, enabling reliable prevention of both known and unseen failures.

III. Setup: Latent Safety Filters via Reachability Analysis in a World Model

In this section, we briefly introduce the latent safety filters [41] for systems with hard-to-model dynamics and safety specifications inferred from high-dimensional observations.

Latent World Model. To model complex systems directly from sensor observations, we train a world model [19] using a fixed offline dataset of robot–environment interactions, $\mathcal{D}_{\text{train}} := \{\{(o_t, a_t, l_t)\}_{t=1}^{T}\}_{i=1}^{N_{\text{train}}} \subset \mathcal{D}_{\text{WM}}$, consisting of trajectories with high-dimensional observations $o \in \mathcal{O}$, robot actions $a \in \mathcal{A}$, and failure labels $l \in \{-1, 1\}$ indicating visible safety hazards. The latent world model consists of an encoder $\mathcal{E}$ that maps an observation into the latent representation $z \in \mathcal{Z}$ and a latent dynamics model:

$$\begin{aligned} \text{Encoder:} \quad & z_t \sim \mathcal{E}(z_t \mid \hat{z}_t, o_t) \\ \text{Dynamics:} \quad & \hat{z}_t \sim f_z(\hat{z}_t \mid z_{t-1}, a_{t-1}). \end{aligned} \tag{1}$$

Safety Specification ($\mathcal{F}$). Hard-to-model safety constraints (e.g., spilling, block toppling) are specified in the latent space via a failure set $\mathcal{F} := \{z : \ell_z(z) \le 0\} \subset \mathcal{Z}$ encoded via the zero-sublevel set of a margin function ℓ_z. In practice, ℓ_z is a binary classifier $l_t = \ell_z(z_t)$ learned with $\mathcal{D}_{\text{train}}$.

Computing Latent Safety Filters (π^{π}, V^{π}). Following [41], we conduct HJ reachability analysis [39, 24] in the latent space to synthesize both a safety value function $V^{\pi} : \mathcal{Z} \to \mathbb{R}$ and a safety-preserving policy $\pi^{\pi} : \mathcal{Z} \to \mathcal{A}$, entirely within the imagination of the world model. Specifically, we solve the fixed-point safety Bellman equation with a time discounting factor $\gamma \in [0, 1)$ [15]:

$$\begin{aligned} V^{\pi}(z_t) &= (1-\gamma)\ell_z(z_t) + \gamma \min\left\{\ell_z(z_t), \max_{a_t \in \mathcal{A}} V^{\pi}(\hat{z}_{t+1})\right\}, \\ \pi^{\pi}(z_t) &= \arg\max_{a_t \in \mathcal{A}} V^{\pi}(\hat{z}_{t+1}), \quad \hat{z}_{t+1} \sim f_z(z_t, \pi^{\text{task}}). \end{aligned} \tag{2}$$

Intuitively, V^{π} represents how close the robot comes to failure starting from z_t despite its best efforts, and π^{π} is a maximally safety-preserving policy. Note that, in contrast to typical RL for reward maximization, this optimization performs a *min-over-time* to *remember* safety-critical events. Therefore, $V^{\pi} < 0$ indicates that the robot is doomed to fail, while $V^{\pi} \ge 0$ means that there exists a safety-preserving action to prevent failures (e.g., returned by π^{π}).

Runtime Safety Filtering. At runtime, the latent safety filter safeguards an arbitrary task policy π^{task} based on the current observations and proposed action. By checking V^{π} as a monitor with a small margin $\delta \approx 0$, the safety filter either allows π^{task} or overrides it with the fallback policy π^{π}:

$$a^{\text{exec}} := \mathbb{1}\{\underbrace{V^{\pi}(z') > \delta}_{\pi^{\text{task}} \text{ is safe, proceed}}\}\pi^{\text{task}} + \mathbb{1}\{\underbrace{V^{\pi}(z') \le \delta}_{\pi^{\text{task}} \text{ is unsafe}}\}\pi^{\pi}(z), \tag{3}$$

where the value function is evaluated at the next latent state predicted by the learned world model $z' \sim f_z(z, \pi^{\text{task}})$.

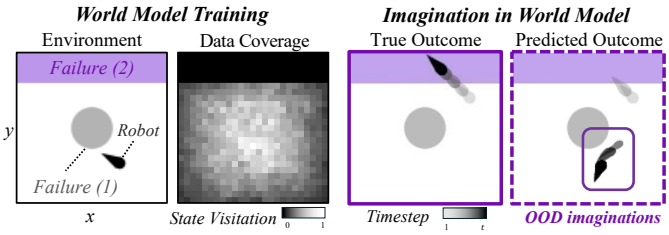

Fig. 2: WM imaginations can lead to ***OOD Failures***.

Challenge: Unreliable World Model Predictions Can Result in OOD Failures. While latent safety filters can compute control strategies that prevent hard-to-model failures, their training (2) and runtime filtering (3) rely on imagined futures generated by the latent dynamics model. However, a pretrained world model can hallucinate in uncertain scenarios where it lacks knowledge, leading to *OOD failures*.

Consider the simple example in Fig. 2 where a Dubins car must avoid two failure sets: a circular grey and a rectangular purple region. The world model is trained with RGB images of the environment and angular velocity actions, but the model training data lacks knowledge of the robot entering the purple failure set. When the world model imagines an action sequence in which the robot enters this region (*third* image of Fig. 2), the world model hallucinates as soon as the scenario goes out-of-distribution: the robot teleports away from the failure region and to a safe state (*rightmost* image of Fig. 2). This phenomenon leads to latent safety filters that cannot prevent unseen failures, and even known failures, due to optimistic safety estimates of uncertain out-of-distribution scenarios.

IV. Uncertainty-aware Latent Safety Filters

To formalize reliable safe control in latent space, our key idea is to use the epistemic uncertainty of the world model as a proxy for detecting safety hazards not represented in the training dataset. Specifically, we augment the safety specification that accounts for ***known failures***–scenarios the world model can anticipate with confidence–with ***OOD failures***: potentially unsafe, out-of-distribution scenarios where the model's imaginations are highly uncertain and lose their reliability.

Uncertainty-aware Latent Space & Dynamics. We quantify the epistemic uncertainty of the world model, $u \in \mathbb{R}$, to identify OOD imaginations of the world model. To assess the reliability of latent dynamics predictions, the uncertainty should capture the *dynamics uncertainty* induced by latent–action transitions (z, a). This is crucial because generative world models are prone to hallucination, often producing in-distribution predictions when exposed to OOD inputs. Therefore, OOD detection methods that rely solely on the predicted latent state z are overconfident, as predicted latents from OOD scenarios are projected into in-distribution representations, as depicted in Fig. 2.

We then augment the latent space to incorporate this epistemic uncertainty, $\tilde{z}_t = (z_t, u_t)^\top \in \mathcal{Z} \times \mathbb{R}$. This formulation enables modeling both *known failures* $\mathcal{F}_{\text{known}} := \{\tilde{z} \mid \ell_z(z) < 0\}$, which are predictable with the learned model, and *OOD failures* $\mathcal{F}_{\text{OOD}} := \{\tilde{z} \mid u > \epsilon\}$ which are OOD imaginations with quantified uncertainty exceeding a predefined threshold ϵ. The latent dynamics and safety margin function are extended to operate in the augmented latent space:

$$f_{\tilde{z}}(\tilde{z}_{t+1} \mid \tilde{z}_t, a_t) = \left[f_z(z_{t+1} \mid z_t, a_t),\ \mathrm{D}(z_t, a_t)\right]^\top, \tag{4}$$

$$\ell_{\tilde{z}}(\tilde{z}_t) = \min\left\{\ell_z(z_t), \kappa\left(\epsilon - u_t\right)\right\}, \tag{5}$$

with $\kappa \in \mathbb{R}^+$. The uncertainty $u_{t+1} = \mathrm{D}(z_t, a_t)$ is obtained via measuring reliability of a transition (described in Sec V-A) and the uncertainty-aware failure set $\tilde{\mathcal{F}}$ is represented via the zero sub-level set of the augmented margin function: $\tilde{\mathcal{F}} := \mathcal{F}_{\text{known}} \cup \mathcal{F}_{\text{OOD}} = \{\tilde{z} \mid \ell_{\tilde{z}}(\tilde{z}) < 0\}$.

Uncertainty-aware Latent Reachability Analysis. We compute a latent safety filter via Eq. 2 and perform safety filtering as in Eq. 3, but use the uncertainty-aware latent dynamics from Eq. 4 throughout. This formulation ensures that the value function assigns negative values to OOD scenarios where the uncertainty exceeds a predefined threshold. By explicitly penalizing such transitions, the resulting safety filter discourages the system from entering OOD regions while also avoiding known, predictable failures. This mitigates overly optimistic imaginations and enables the filter to reliably learn both a safety monitor and a fallback policy that proposes safe, in-distribution actions.

V. Computing Uncertainty-aware Latent Safety Filters

While the prior section formalized the uncertainty-aware latent safety filter by augmenting the latent space with the world model's epistemic uncertainty, we face two key challenges when instantiating our the uncertainty-aware latent safety filter in practice: (i) How can we quantify the epistemic uncertainty of the world model? (Sec. V-A) and (ii) How can we ensure the OOD threshold ϵ is *calibrated* to reliably detect OOD failures based on quantified epistemic uncertainty? (Sec. V-B).

A. Quantifying Epistemic Uncertainty in World Models

Training a Probabilistic Ensemble Latent Predictor. To capture the epistemic uncertainty of the world model, we employ an ensemble of next-latent predictors, $E := \{\hat{f}_z^k\}_{k=1}^K$, which is a separate module regressing the pretrained latent dynamics f_z. Each ensemble member is initialized with distinct parameters ψ_k and trained to predict the next latent z_{t+1} given the current latent z_t and action a_t with Gaussian negative log-likelihood loss:

$$\begin{aligned} &\text{Latent Predictor:} \quad z_{t+1}^k \sim \hat{f}_z^k(z_t, a_t; \psi_k), \\ &\text{where} \quad \hat{f}_z^k(z_t, a_t; \psi_k) := \mathcal{N}(\mu_{\psi_k}(z_t, a_t), \Sigma_{\psi_k}(z_t, a_t)). \end{aligned} \tag{6}$$

μ_{ψ_k} and Σ_{ψ_k} denote the predicted mean and diagonal covariance, respectively. Note that the covariance models the *aleatoric uncertainty* inherent in the latent dynamics due to partial observability and stochasticity. The ensemble latent predictor is trained on latent transitions $\{\{(z_t, a_t, z_{t+1})\}_{t=1}^{T-1}\}_{i=1}^{N_{\text{train}}}$, encoded from a pretrained latent world model with the same offline dataset $\mathcal{D}_{\text{train}}$ used for world model training.

Epistemic Uncertainty Quantification. While empirical variance over ensemble predictions is widely used as an uncertainty measure [29, 71, 52, 64], this conflates aleatoric uncertainty (i.e., inherent uncertainty of latent dynamics) with epistemic uncertainty (i.e., uncertainty arising from a lack of knowledge). Since our goal is to control away from *OOD failures*, which the world model has never encountered and thus cannot reliably predict, it is essential to focus explicitly on the model's *epistemic uncertainty* to form our constraint on the latent space. Otherwise, the safety filter may fail to reject unsafe OOD imaginations or become overly conservative in response to intrinsic stochasticity. Following [54, 30], we quantify the epistemic uncertainty of the latent dynamics $\mathrm{D}(z_t, a_t)$ via the Jensen-Rényi Divergence (JRD) [45] of the ensemble predictions:

$$\overbrace{u_{t+1} = \mathrm{D}(z_t, a_t)}^{\text{Epistemic Uncertainty}} := \overbrace{H_\alpha\Big(\sum_{k=1}^{K} \tfrac{1}{K} \hat{f}_z^k\Big)}^{\text{Total Uncertainty}} - \overbrace{\sum_{k=1}^{K} \tfrac{1}{K} H_\alpha(\hat{f}_z^k)}^{\text{Aleatoric Uncertainty}}, \tag{7}$$

where $H_\alpha(Z)$ is Rényi entropy with a random variable Z:

$$H_\alpha(Z) = \frac{1}{1-\alpha} \log \int p(z)^\alpha \, \mathrm{d}z. \tag{8}$$

B. Detecting OOD Imaginations via Conformal Prediction

Recall that during reachability analysis, *OOD failures* are detected when the uncertainty of an imagined transition exceeds a threshold, $\mathcal{F}_{\text{OOD}} = \{\tilde{z} \mid u > \epsilon\}$. However, setting this threshold is nontrivial: too strict a threshold can result in high false-positive rates (misclassifying in-distribution transitions as OOD), leading to overly conservative filters; too loose a threshold may fail to detect true OOD transitions. We employ conformal prediction (CP) [62, 4] to automatically calibrate the threshold $\epsilon \in \mathbb{R}$ in a principled way, using a held-out calibration dataset $\mathcal{D}_{\text{calib}} = \mathcal{D}_{\text{WM}} \setminus \mathcal{D}_{\text{train}}$.

In-distribution Recall Guarantee via Class-Conditioned Conformal Prediction. CP typically requires the calibration set $\mathcal{D}_{\text{calib}}$ to contain both inputs to the prediction model (e.g., (z_t, a_t)) and their corresponding ground-truth labels (e.g., ID or OOD). Unfortunately, in our setting, true OOD labels are, by definition, not accessible. As such, we assume the calibration dataset consists only of in-distribution transitions. Formally, we adopt class-conditioned conformal prediction [12, 9] to calibrate the uncertainty threshold ϵ, providing conditional recall guarantees for detecting in-distribution transitions with user-defined confidence level $\alpha_{\text{cal}} \in [0, 1]$:

$$\mathbb{P}\left(\mathrm{D}(z_t, a_t) < \hat{\epsilon} \mid (z_t, a_t) \in \mathcal{D}_{\text{WM}}\right) \geq 1 - \alpha_{\text{cal}}, \tag{9}$$

Intuitively, conformal prediction can help us select an uncertainty threshold $\hat{\epsilon}$ such that in-distribution latent transitions can be detected with probability at least $1 - \alpha_{\text{cal}}$. Conversely, latent transitions with uncertainty greater than this threshold can be interpreted as OOD.

Trajectory-Level Calibration. While standard class-conditioned conformal prediction assumes exchangeability of the data, this assumption does not hold in our setting, as each transition depends on the full history of latent states and actions. To address this, we adopt a trajectory-level calibration approach [44], assuming that the calibration trajectories $\tau_i = \{(z_t, a_t)\}_{t=1}^{T} \in \mathcal{D}_{\text{calib}}$ are drawn i.i.d. from the same distribution as the world model training data, $\{\tau_i\}_{i=1}^{N} \overset{\text{iid}}{\sim} \mathcal{D}_{\text{WM}}$. For each trajectory, we define the trajectory-level nonconformity score $Q_{\tau_i}^{\alpha_{\text{trans}}}$ as the $(1-\alpha_{\text{trans}})$-quantile of the set of quantified epistemic uncertainties $\{u_t\}_{t=1}^{T}$. This ensures that at most an α_{trans} fraction of a trajectory's uncertainty values exceed $Q_{\tau_i}^{\alpha_{\text{trans}}}$, making the estimate more robust to noise in uncertainty predictions. We then determine the calibration threshold $\hat{\epsilon}$ as the $(1 - \alpha_{\text{cal}})$-quantile of the set $\{Q_{\tau_i}^{\alpha_{\text{trans}}}\}_{i=1}^{N}$ by selecting the $\lceil (1 - \alpha_{\text{cal}})(N + 1) \rceil$-th smallest value over trajectories. With the exchangeability assumption between calibration and test trajectories, conformal prediction guarantees that for a new test trajectory $\tau_{\text{test}} = \{(z_t^{\text{test}}, a_t^{\text{test}})\}_{t=1}^{T}$, the following probabilistic guarantee holds:

$$\mathbb{P}_{\tau_{\text{test}} \sim \mathcal{D}_{\text{WM}}}\left(\mathbb{P}_t\left\{\mathrm{D}(z_t^{\text{test}}, a_t^{\text{test}}) \leq \hat{\epsilon}\right\} \geq 1 - \alpha_{\text{trans}}\right) \geq 1 - \alpha_{\text{cal}}. \tag{10}$$

Although this guarantee applies only to in-distribution data, it ensures a low false positive rate by bounding the probability of misclassifying in-distribution transitions as OOD. Specifically, the probability that the trajectory-level nonconformity score exceeds the threshold for in-distribution data is bounded by $\mathbb{P}_{\tau_{\text{test}} \sim \mathcal{D}_{\text{WM}}}\left(Q_{\tau_{\text{test}}}^{\alpha_{\text{trans}}} \geq \hat{\epsilon}\right) \leq \alpha_{\text{cal}}$. As a result, any transition with a quantified epistemic uncertainty above $\hat{\epsilon}$ can be reliably classified as OOD, since such events are guaranteed to be rare under the in-distribution distribution.

VI. Simulation & Hardware Experiments

A. A Benchmark Safe Control Task with a 3D Dubins Car

We first conduct experiments with a simple benchmark safe navigation task where privileged information about the state, dynamics, safe set, and safety controller is available. The world model is trained solely from image observations of the environment, without access to privileged state information.

Privileged Dynamics: Dubins Car. Let the privileged Dubins car state be $s = [p_x, p_y, \theta]$, with discrete-time dynamics $s_{t+1} = s_t + \Delta t\,[v\cos(\theta_t),\, v\sin(\theta_t),\, a_t]$. We assume a fixed velocity $v = 1\,\mathrm{m/s}$, time step $\Delta t = 0.05\,\mathrm{s}$, and discrete action space $a_t \in \mathcal{A} = \{-1.25, 0, 1.25\}$ rad/s.

Evaluation & Metrics. Given access to ground-truth dynamics, we compute the ground-truth safety value function using grid-based methods [40], enabling direct evaluation of the safety monitor $V^{\textcolor{black}{\blacktriangledown}}$)'s classification accuracy across all three state dimensions. To assess $\pi^{\textcolor{black}{\blacktriangledown}}$, we roll out the learned policies from safe initial states with positive ground-truth safety values and measure the safety rate by checking whether the resulting trajectories remain safe without violating constraints.

Baselines. We evaluate ***UNISafe***, which learns the uncertainty-aware unsafe set $\tilde{\mathcal{U}}$ from the failure set $\tilde{\mathcal{F}} = \mathcal{F}_{\text{known}} \cup \mathcal{F}_{\text{OOD}}$, against ***LatentSafe***, which considers only known failures, $\mathcal{F}_{\text{known}}$. Also, we compare *JRD* with other

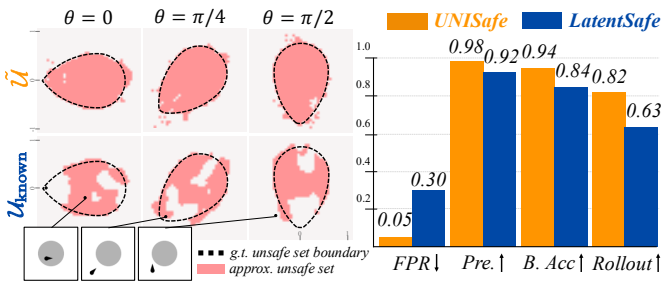

Fig. 3: ***UNISafe*** **vs** ***LatentSafe***. Without *OOD failures*, the safety value learned from the unreliable world model leads to higher FPR, overconfidently classifying unsafe states as safe.

OOD detection baselines to assess uncertainty quantification. *TotalUncertainty* compute variance of mean predictions across the ensemble without isolating aleatoric components [52, 43, 53]. *MaxAleatoric* uses the maximum predicted ensemble variance, $\max_k \|\Sigma_{\psi_k}(z_t, a_t)\|_F$, representing aleatoric uncertainty [71, 57]. *DensityEst* employs neural spline flows [13, 28] to compute likelihoods of (z) or (z, a) for OOD detection. For every method, thresholds are calibrated with the same held-out calibration dataset, and Double DQN (DDQN) [60] is used to train all the safety value functions.

UNISafe **reliably identifies the OOD failure $\mathcal{F}_{\text{OOD}}$.** To evaluate OOD detection, we first consider a setting where failure states are never observed by $\mathcal{D}_{\text{train}}$. The ground-truth failure set is defined as $|p_y| > 0.6$, while the offline dataset contains only 1000 safe trajectories that never enter this region, making the failure set entirely OOD. Table I shows that *JRD* achieves the highest balanced accuracy (B.Acc.) compared to other OOD detection methods, whereas methods not targeting epistemic uncertainty exhibit higher FPRs and lower balanced accuracies. Additionally, *DensityEst* based only on z shows low TNR, highlighting the necessity of latent-action transition-based OOD detection.

	Method	TPR↑	TNR↑	B.Acc.↑
UQ	*TotalUncertainty*	0.88	0.97	0.93
	MaxAleatoric	0.78	0.88	0.83
	DensityEst (z, a)	0.98	0.87	0.92
	DensityEst (z)	0.99	0.56	0.77
Calib	*JRD* $(\epsilon = \hat{\epsilon})$	0.93	0.95	0.94
	JRD $(\epsilon = \hat{\epsilon} + 0.3)$	0.98	0.43	0.71
	JRD $(\epsilon = \hat{\epsilon} - 0.3)$	0.85	0.96	0.90

TABLE I: Safety value function quality with different OOD detection methods.

A calibrated OOD threshold yields a higher quality value function. We perturb our calibrated threshold $\hat{\epsilon}$ to obtain $\epsilon = \hat{\epsilon} \pm 0.3$ and study the sensitivity of the value function to threshold selection. Table I shows that our automatic calibration process selects thresholds that lead to value functions with both high TPR and TNR, unlike the uncalibrated thresholds that degrade accuracy.

UNISafe **robustly learns safety filters despite high uncertainties in the world models.** We evaluate whether our method can synthesize a robust safety filter with uncertain world due to limited data coverage. In this setting, the vehicle must avoid a circular obstacle of radius 0.5 m at center, with the failure set defined as $p_x^2 + p_y^2 < 0.5^2$, and $\mathcal{D}_{\text{train}}$ consists of both safe and unsafe trajectories. We construct a dataset of 1000 expert trajectories that never enter the ground-truth unsafe sets and 50 random trajectories that may include failure states. Expert trajectories are generated using the ground-truth safety value, applying fallback actions near the unsafe boundary and random actions elsewhere, inducing high uncertainty around the unsafe boundary. Fig. 3 shows that ***UNISafe*** robustly learns the safety monitor with higher balanced accuracy, whereas ***LatentSafe*** overconfidently misclassifies unsafe states as safe. In rollouts from 181 challenging safe initial states, where the vehicle is oriented toward failure, ***UNISafe*** also achieves higher safety rates.

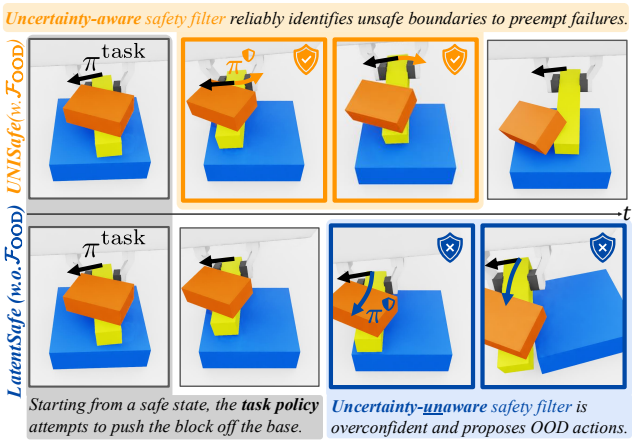

Fig. 4: ***UNISafe*** prevents failure by proposing in-distribution, safe backup actions, while ***LatentSafe*** fails to preempt it by overestimating unsafe OOD actions.

B. Simulation: Vision-Based Block Plucking

Setup. We scale our method to a visual manipulation task using IsaacLab [37], where a Franka manipulator must pluck the middle block from a stack of three while ensuring the top one remains on the bottom one. Observations consist of images from a wrist-mount and a tabletop camera, with 7-D proprioceptive inputs. Actions are a 6-DoF end-effector delta pose with a discrete gripper command.

Evaluations. We adopt DreamerV3 [20] as our task policy π^{task}, trained with a dense reward signal to achieve the task

Method	f_z	$\mathcal{F}_{\text{known}}$	$\mathcal{F}_{\text{OOD}}$	Safe Success (↑)	Failure (↓)	Incompletion	Model Error (↓)
No Filter (π^{task})	-	-	-	0.58	0.41	0.01	59.3 ± 3.3
CQL [32]	✗	✓	✗	0.63	0.33	0.04	50.9 ± 11.5
COMBO [72]	✓	✓	✗	0.47	0.41	0.12	51.6 ± 12.8
SafeOnly	✓	✗	✓	0.71	0.28	0.01	46.9 ± 2.6
LatentSafe [41]	✓	✓	✗	0.68	0.30	0.01	60.2 ± 4.7
UNISafe (*JRD*)	✓	✓	✓	0.72	0.20	0.08	43.1 ± 1.2
(*TotalUncertainty*)	✓	✓	✓	0.54	0.18	0.28	39.1 ± 4.1
(*MaxAleatoric*)	✓	✓	✓	0.64	0.25	0.11	41.4 ± 9.1
(*DensityEst*)	✓	✓	✓	0.66	0.24	0.10	41.4 ± 5.2

TABLE II: *Rollout Results on Block Plucking.* Safe success is plucking a block without failure, and incompletion is a timeout without success or failure. The average world model training loss per trajectory is reported as a proxy for uncertainty.

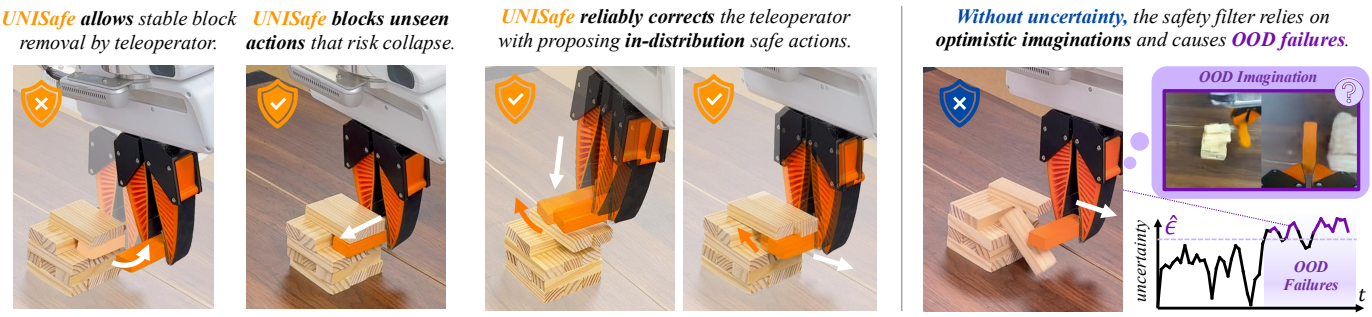

Fig. 5: *Teleoperator Playing Jenga with Safety Filters.* ***UNISafe*** enables non-conservative yet effective filtering of the teleoperator's actions, ensuring the system remains within the in-distribution regions. In contrast, the uncertainty-unaware safety filter ***LatentSafe*** optimistically treats uncertain actions as safe, leading to failure.

with a soft penalty for failures. The training dataset $\mathcal{D}_{\text{train}}$ consists of 3000 trajectories comprising both safe and unsafe behavior rolled out from π^{task}. We adopt Soft Actor-Critic (SAC) [18] as our solver for latent reachability. For evaluation, task policy rollouts are filtered using the safety filter with $\delta = 0.1$, evaluated over 1000 random initial conditions.

Baselines. As in Sec. VI-A, we compare ***UNISafe*** with ***LatentSafe*** trained on the same dataset with and without $\mathcal{F}_{\text{OOD}}$, as well as different OOD detection baselines. ***SafeOnly*** learns a WM and latent safety filter only on successful demonstrations without $\mathcal{F}_{\text{known}}$, implicitly treating all failures as $\mathcal{F}_{\text{OOD}}$, as in [28, 8, 68]. Also, we adapt *CQL* [32] and *COMBO* [72] to optimize Eq. 2 with conservative losses, but without uncertainty quantification.

UNISafe* minimizes failure by preventing safety overestimation.** Table II shows that ***UNISafe, which incorporates both known and OOD failures, achieves the lowest failure rates and model errors. In contrast, ***LatentSafe*** overestimates the safety of OOD actions, leading to unsafe action proposals, as shown in Fig. 4. ***SafeOnly*** shows limited effectiveness, showing OOD detection from success-only data is insufficient in complex settings. Offline RL with conservative losses performs even worse than ***LatentSafe***, indicating that conservatism alone cannot replace failure set identification.

Quantifying epistemic uncertainty leads to safe but non-conservative behaviors. While all OOD detection methods improve filtering performance over ***LatentSafe***, targeting aleatoric uncertainty (*TotalUncertainty* and *MaxAleatoric*) tends to be overly conservative, resulting in higher incompletion rates and more frequent interventions. In contrast, ***UNISafe*** with *JRD* explicitly targets epistemic uncertainty and achieves the most reliable performance. *DensityEst* shows limited performance, highlighting the challenge of modeling likelihood in high-dimensional latent spaces.

C. Hardware: Vision-based Jenga with a Robotic Manipulator

Setup. We evaluate our method on a real-world robotic manipulation task using a fixed-base Franka Research 3 arm, equipped with a third-person camera and a wrist-mounted camera. The robot must extract a target block from a tower without collapsing, then place it on top. For $\mathcal{D}_{\text{train}}$, we collect 720 trajectories: 150 random (no contact), 480 successful, and 90 failure cases.

UNISafe* reliably filters both known and unseen failures.** First, a teleoperator is π^{task}, controlling the end-effector pose and gripper while assisted by ***UNISafe. As shown in Fig. 5, the teleoperator can freely execute safe behaviors, which require careful tilting and precise block manipulation that are non-trivial to perform. When erratic or OOD actions are attempted, posing a risk of tower collapse, ***UNISafe*** reliably intervenes to correct the behavior and maintain stability within the in-distribution region. In contrast, ***LatentSafe*** fails to preemptively detect such boundaries due to optimistic OOD imagination, ultimately allowing high-uncertainty actions. Next, we quantitatively evaluate filtering by replaying 50 failure trajectories as π^{task} that result in tower collapse. The corresponding action sequences are replayed as a task policy with either ***UNISafe*** or ***LatentSafe*** as the safety filter. Fig. 6 shows that ***UNISafe*** leads to lower failure rates and maintains low model uncertainty.

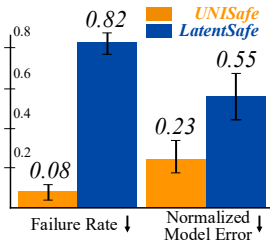

Fig. 6: Filtering π^{task} on hardware.

VII. Conclusion

In this work, we propose ***UNISafe***, a framework for reliable latent-space safe control that *unifies* reachability analysis in a latent world model with OOD detection of the world model predictions. To detect unreliable out-of-distribution imaginations of the world model, we introduce a principled method to quantify the world model's epistemic uncertainty and calibrate a threshold. We then augment the latent space with epistemic uncertainty and perform an uncertainty-aware latent reachability analysis to synthesize a safety filter that reliably safeguards arbitrary policies from both known failures and unseen safety hazards. We demonstrate that our approach reliably identifies OOD imaginations and synthesizes an uncertainty-aware latent safety filter from an offline dataset with limited coverage, enabling safe control in complex vision-based tasks by preemptively detecting safety risks and proposing safe, in-distribution backup actions.

LIMITATIONS

Component vs. System-level Safety Assurances. While our uncertainty-aware safety filter empirically can prevent both seen and unseen failures by incorporating OOD failures, it does not formally guarantee zero failure rates. In this work, we only provide a *component-level* statistical assurance on detecting OOD transitions within the world model via conformal prediction. Future work should study *system-level* assurances on the overall safety filter that is also influenced by our RL approximations in high-dimensional learned latent spaces. Moreover, our framework assumes that the system starts from an in-distribution safe initial state and that no unknown disturbances or visual distractions appear during operation. For robust deployment, a system-level failure monitoring mechanism is necessary, which can reliably detect when the system loses its confidence. While our supplementary experiments indicate that our uncertainty measure can be leveraged for such system-level failure detection (see Appendix. D), further exploration on system-level failure detection and mitigation remains as an important future work [55, 2].

Limited Generalizability and Reliability. Our latent safety filter relies on the capabilities of the learned world model. While recent generative world models have demonstrated promising results [73, 1], the world model's predictions can be imprecise even within in-distribution regions or fail to generalize to unseen scenarios. Although our safety filter adopts a minimally conservative approach to uncertain scenarios, its performance can be further improved with additional data. Future work should explore safe exploration strategies or active learning methods, using quantified epistemic uncertainty as intrinsic rewards to enhance world model generalization.

Challenges in Uncertainty Quantification. While our method adopts epistemic uncertainty quantification as a proxy for detecting unreliable world model imaginations, there are several limitations to this approach. Even within regions that are nominally in-distribution, world model predictions can still be imprecise or biased, particularly in complex or stochastic systems. In other words, while a transition may be classified as in-distribution, this does not guarantee the correctness of the model's prediction, potentially leading to an imprecise safety filter. Moreover, our uncertainty quantification assumes a Gaussian distribution over the next latent prediction, which may not hold in systems with complex, multimodal dynamics. It also adopts an ensemble as a separate module from the world model, which may not faithfully capture the model's true uncertainty. Exploring methods for faithfully detecting OOD scenarios under complex, multimodal data distributions presents an important direction for future work. Additionally, our framework and the safety Bellman equation (2) does not account for aleatoric uncertainty, and thus optimizes for the expected safety violation. Extending the framework to explicitly model aleatoric uncertainty in the latent dynamics could improve robustness, enabling latent-space safe control that better anticipates worst-case outcomes under the world model's predictions [70, 16, 47, 42].

REFERENCES

[1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

[2] Christopher Agia, Rohan Sinha, Jingyun Yang, Zi-ang Cao, Rika Antonova, Marco Pavone, and Jeannette Bohg. Unpacking failure modes of generative policies: Runtime monitoring of consistency and progress. In *Conference on Robot Learning (CoRL)*, 2024.

[3] Aaron D Ames, Xiangru Xu, Jessy W Grizzle, and Paulo Tabuada. Control barrier function based quadratic programs for safety critical systems. *IEEE Transactions on Automatic Control*, 62(8):3861–3876, 2016.

[4] Anastasios N Angelopoulos, Stephen Bates, et al. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023.

[5] Chenjia Bai, Lingxiao Wang, Zhuoran Yang, Zhi-Hong Deng, Animesh Garg, Peng Liu, and Zhaoran Wang. Pessimistic bootstrapping for uncertainty-driven offline reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2022.

[6] Somil Bansal and Claire J Tomlin. Deepreach: A deep learning approach to high-dimensional reachability. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1817–1824, 2021.

[7] 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.

[8] Fernando Castaneda, Haruki Nishimura, Rowan Thomas McAllister, Koushil Sreenath, and Adrien Gaidon. In-distribution barrier functions: Self-supervised policy filters that avoid out-of-distribution states. In *Learning for Dynamics and Control Conference (L4DC)*, pages 286–299, 2023.

[9] Kaustav Chakraborty, Aryaman Gupta, and Somil Bansal. Enhancing safety and robustness of vision-based controllers via reachability analysis. *arXiv preprint arXiv:2410.21736*, 2024.

[10] 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.

[11] Yusuf Umut Ciftci, Darren Chiu, Zeyuan Feng, Gaurav S Sukhatme, and Somil Bansal. Safe-gil: Safety guided imitation learning for robotic systems. *arXiv preprint arXiv:2404.05249*, 2024.

[12] Tiffany Ding, Anastasios Angelopoulos, Stephen Bates, Michael Jordan, and Ryan J Tibshirani. Class-conditional conformal prediction with many classes. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:

64555–64576, 2023.

[13] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

[14] Jaime F. Fisac, Anayo K. Akametalu, Melanie N. Zeilinger, Shahab Kaynama, Jeremy Gillula, and Claire J. Tomlin. A general safety framework for learning-based control in uncertain robotic systems. *IEEE Transactions on Automatic Control*, 64(7):2737–2752, 2018.

[15] Jaime F Fisac, Neil F Lugovoy, Vicenç Rubies-Royo, Shromona Ghosh, and Claire J Tomlin. Bridging hamilton-jacobi safety analysis and reinforcement learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 8550–8556, 2019.

[16] Milan Ganai, Zheng Gong, Chenning Yu, Sylvia Herbert, and Sicun Gao. Iterative reachability estimation for safe reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:69764–69797, 2023.

[17] Milan Ganai, Sicun Gao, and Sylvia Herbert. Hamilton-jacobi reachability in reinforcement learning: A survey. *IEEE Open Journal of Control Systems*, 2024.

[18] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning (ICML)*, pages 1861–1870, 2018.

[19] 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)*, pages 2555–2565. PMLR, 2019.

[20] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, April 2025. ISSN 1476-4687.

[21] Tairan He, Chong Zhang, Wenli Xiao, Guanqi He, Changliu Liu, and Guanya Shi. Agile but safe: Learning collision-free high-speed legged locomotion. In *Robotics: Science and Systems (RSS)*, 2024.

[22] Sylvia Herbert, Jason J Choi, Suvansh Sanjeev, Marsalis Gibson, Koushil Sreenath, and Claire J Tomlin. Scalable learning of safety guarantees for autonomous systems using hamilton-jacobi reachability. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 5914–5920, 2021.

[23] Kai-Chieh Hsu, Vicenç Rubies-Royo, Claire J. Tomlin, and Jaime F. Fisac. Safety and liveness guarantees through reach-avoid reinforcement learning. In *Robotics: Science and Systems (RSS)*, 2021.

[24] Kai-Chieh Hsu, Haimin Hu, and Jaime F Fisac. The safety filter: A unified view of safety-critical control in autonomous systems. *Annual Review of Control, Robotics, and Autonomous Systems*, 7, 2023.

[25] Kai-Chieh Hsu, Duy Phuong Nguyen, and Jaime Fernandez Fisac. Isaacs: Iterative soft adversarial actor-critic for safety. In *Learning for Dynamics and Control Conference (L4DC)*, pages 90–103, 2023.

[26] Kai-Chieh Hsu, Allen Z Ren, Duy P Nguyen, Anirudha Majumdar, and Jaime F Fisac. Sim-to-lab-to-real: Safe reinforcement learning with shielding and generalization guarantees. *Artificial Intelligence*, 314:103811, 2023.

[27] Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning (ICML)*, pages 5084–5096, 2021.

[28] Katie Kang, Paula Gradu, Jason J Choi, Michael Janner, Claire Tomlin, and Sergey Levine. Lyapunov density models: Constraining distribution shift in learning-based control. In *International Conference on Machine Learning (ICML)*, pages 10708–10733, 2022.

[29] Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:21810–21823, 2020.

[30] Taekyung Kim, Jungwi Mun, Junwon Seo, Beomsu Kim, and Seongil Hong. Bridging active exploration and uncertainty-aware deployment using probabilistic ensemble neural network dynamics. In *Robotics: Science and Systems (RSS)*, 2023.

[31] Victor Kolev, Rafael Rafailov, Kyle Hatch, Jiajun Wu, and Chelsea Finn. Efficient imitation learning with conservative world models. In *Learning for Dynamics and Control Conference (L4DC)*, pages 1777–1790, 2024.

[32] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:1179–1191, 2020.

[33] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

[34] Lars Lindemann, Alexander Robey, Lejun Jiang, Satyajeet Das, Stephen Tu, and Nikolai Matni. Learning robust output control barrier functions from safe expert demonstrations. *IEEE Open Journal of Control Systems*, 2024.

[35] Huihan Liu, Shivin Dass, Roberto Martín-Martín, and Yuke Zhu. Model-based runtime monitoring with interactive imitation learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 4154–4161, 2024.

[36] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems (NeurIPS)*, 33: 21464–21475, 2020.

[37] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, and Gavriel State. Isaac gym: High performance gpu-based physics simulation for robot learning, 2021.

[38] Russell Mendonca, Oleh Rybkin, Kostas Daniilidis, Danijar Hafner, and Deepak Pathak. Discovering and

achieving goals via world models. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:24379–24391, 2021.

[39] Ian M Mitchell, Alexandre M Bayen, and Claire J Tomlin. A time-dependent hamilton-jacobi formulation of reachable sets for continuous dynamic games. *IEEE Transactions on Automatic Control*, 50(7):947–957, 2005.

[40] Ian M Mitchell et al. A toolbox of level set methods. *UBC Department of Computer Science Technical Report TR-2007-11*, 1:6, 2007.

[41] Kensuke Nakamura, Lasse Peters, and Andrea Bajcsy. Generalizing safety beyond collision-avoidance via latent-space reachability analysis. *Robotics: Science and Systems (RSS)*, 2025.

[42] Duy P. Nguyen, Kai-Chieh Hsu, Wenhao Yu, Jie Tan, and Jaime Fernández Fisac. Gameplay filters: Robust zero-shot safety through adversarial imagination. In *Conference on Robot Learning (CoRL)*, 2024.

[43] Rafael Rafailov, Tianhe Yu, Aravind Rajeswaran, and Chelsea Finn. Offline reinforcement learning from images with latent space models. In *Learning for Dynamics and Control Conference (L4DC)*, pages 1154–1168, 2021.

[44] Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. Robots that ask for help: Uncertainty alignment for large language model planners. In *Conference on Robot Learning (CoRL)*, 2023.

[45] Alfréd Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 4, pages 547–562. University of California Press, 1961.

[46] Charles Richter and Nicholas Roy. Safe visual navigation via deep learning and novelty detection. In *Robotics: Science and Systems (RSS)*, 2017.

[47] Marc Rigter, Bruno Lacerda, and Nick Hawes. RAMBO-RL: Robust adversarial model-based offline reinforcement learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[48] Alexander Robey, Haimin Hu, Lars Lindemann, Hanwen Zhang, Dimos V Dimarogonas, Stephen Tu, and Nikolai Matni. Learning control barrier functions from expert demonstrations. In *IEEE Conference on Decision and Control (CDC)*, pages 3717–3724, 2020.

[49] Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021.

[50] Mohammadreza Salehi, Hossein Mirzaei, Dan Hendrycks, Yixuan Li, Mohammad Hossein Rohban, and Mohammad Sabokrou. A unified survey on anomaly, novelty, open-set, and out of-distribution detection: Solutions and future challenges. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.

[51] Robin Schmid, Deegan Atha, Frederik Schöller, Sharmita Dey, Seyed Fakoorian, Kyohei Otsu, Barry Ridge, Marko Bjelonic, Lorenz Wellhausen, Marco Hutter, et al. Self-supervised traversability prediction by learning to reconstruct safe terrain. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12419–12425, 2022.

[52] 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)*, pages 8583–8592, 2020.

[53] Tim Seyde, Wilko Schwarting, Sertac Karaman, and Daniela Rus. Learning to plan optimistically: Uncertainty-guided deep exploration via latent model ensembles. In *Conference on Robot Learning (CoRL)*, pages 1156–1167, 2022.

[54] Pranav Shyam, Wojciech Jaśkowski, and Faustino Gomez. Model-based active exploration. In *International Conference on machine learning (ICML)*, pages 5779–5788, 2019.

[55] Rohan Sinha, Apoorva Sharma, Somrita Banerjee, Thomas Lew, Rachel Luo, Spencer M Richards, Yixiao Sun, Edward Schmerling, and Marco Pavone. A system-level view on out-of-distribution data in robotics. *arXiv preprint arXiv:2212.14020*, 2022.

[56] Rohan Sinha, Amine Elhafsi, Christopher Agia, Matt Foutter, Edward Schmerling, and Marco Pavone. Real-time anomaly detection and reactive planning with large language models. In *Robotics: Science and Systems (RSS)*, 2024.

[57] Yihao Sun, Jiaji Zhang, Chengxing Jia, Haoxin Lin, Junyin Ye, and Yang Yu. Model-bellman inconsistency for model-based offline reinforcement learning. In *International Conference on Machine Learning (ICML)*, pages 33177–33194, 2023.

[58] Brijen Thananjeyan, Ashwin Balakrishna, Suraj Nair, Michael Luo, Krishnan Srinivasan, Minho Hwang, Joseph E. Gonzalez, Julian Ibarz, Chelsea Finn, and Ken Goldberg. Recovery rl: Safe reinforcement learning with learned recovery zones. *IEEE Robotics and Automation Letters*, 6(3):4915–4922, 2021.

[59] Núria Armengol Urpí, Sebastian Curi, and Andreas Krause. Risk-averse offline reinforcement learning. *International Conference on Learning Representations (ICLR)*, 2021.

[60] Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *AAAI Conference on Artificial Intelligence*, 2016.

[61] Marin Vlastelica, Sebastian Blaes, Cristina Pinneri, and Georg Martius. Risk-averse zero-order trajectory optimization. In *Conference on Robot Learning (CoRL)*,

2021.

[62] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

[63] Kim P Wabersich, Andrew J Taylor, Jason J Choi, Koushil Sreenath, Claire J Tomlin, Aaron D Ames, and Melanie N Zeilinger. Data-driven safety filters: Hamilton-jacobi reachability, control barrier functions, and predictive methods for uncertain systems. *IEEE Control Systems Magazine*, 43(5):137–177, 2023.

[64] Hao Wang, Javier Borquez, and Somil Bansal. Providing safety assurances for systems with unknown dynamics. *IEEE Control Systems Letters*, 2024.

[65] Lorenz Wellhausen, René Ranftl, and Marco Hutter. Safe robot navigation via multi-modal anomaly detection. *IEEE Robotics and Automation Letters*, 5(2):1326–1333, 2020.

[66] Albert Wilcox, Ashwin Balakrishna, Brijen Thananjeyan, Joseph E Gonzalez, and Ken Goldberg. Ls3: Latent space safe sets for long-horizon visuomotor control of sparse reward iterative tasks. In *Conference on Robot Learning (CoRL)*, pages 959–969, 2022.

[67] Josiah Wong, Albert Tung, Andrey Kurenkov, Ajay Mandlekar, Li Fei-Fei, Silvio Savarese, and Roberto Martín-Martín. Error-aware imitation learning from teleoperation data for mobile manipulation. In *Conference on Robot Learning (CoRL)*, pages 1367–1378, 2022.

[68] Chen Xu, Tony Khuong Nguyen, Emma Dixon, Christopher Rodriguez, Patrick Miller, Robert Lee, Paarth Shah, Rares Ambrus, Haruki Nishimura, and Masha Itkina. Can we detect failures without failure data? uncertainty-aware runtime failure detection for imitation learning policies. *arXiv preprint arXiv:2503.08558*, 2025.

[69] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024.

[70] 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.

[71] 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.

[72] Tianhe Yu, Aviral Kumar, Rafael Rafailov, Aravind Rajeswaran, Sergey Levine, and Chelsea Finn. Combo: Conservative offline model-based policy optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:28954–28967, 2021.

[73] 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)*, 2025.

APPENDIX

A. A Brief Background on Offline Reinforcement Learning

Offline reinforcement learning (RL) learns policies from a static dataset of past interactions, making it well-suited for applications where online exploration poses safety risks [33, 58, 59, 31]. A major challenge in offline RL is the distribution shift between the learned policy and the behavior policy that collected the data [27, 57], which often leads to overestimation of policy evaluations on OOD scenarios [5]. To address this, conservatism is introduced by penalizing value functions, preventing over-optimism on OOD actions [32, 72, 47]. In offline model-based RL (MBRL), a dynamics model is learned from the static dataset and used to generate synthetic data for policy learning [10, 7, 71, 29, 72, 57]. By quantifying the uncertainty of the learned dynamics model, these methods mitigate model exploitation and discourage the system from entering OOD scenarios. Inspired by this, we quantify uncertainty in a latent dynamics model and ensure a safety filter to proactively prevent the system from entering OOD regions.

B. A Brief Background on HJ Reachability

Hamilton-Jacobi (HJ) reachability is a control-theoretic framework for safety analysis that identifies when current actions may lead to future failures and computes best-effort policies to mitigate such outcomes [39, 24]. Given a dynamical system with state $s \in \mathcal{S}$, action $a \in \mathcal{A}$, and dynamics $s_{t+1} = f(s_t, a_t)$, HJ reachability finds the safe set that can prevent the system from entering a designated *failure set* $\mathcal{F} = \{s \mid \ell(s) < 0\}$, which is represented by a margin function $\ell : \mathcal{S} \to \mathbb{R}$. The framework aims to find the *unsafe set*, denoted $\mathcal{U} \subset \mathcal{S}$, which includes all states from which the system is inevitably driven into $\mathcal{F}$ despite the best effort, and the best effort safety-preserving policy to avoid entering the unsafe set.

The framework jointly computes (i) a safety value $V^{\shield} : \mathcal{S} \to \mathbb{R}$, which quantifies the minimal safety margin the system can achieve from a given state s under optimal behavior, and (ii) a best-effort safety-preserving policy $\pi^{\shield} : \mathcal{S} \to \mathcal{A}$. These are obtained by solving an optimal control problem with the following fixed-point safety Bellman equation:

$$\begin{aligned} V(s) &= \min\Big\{\ell(s),\ \max_{a\in\mathcal{A}} V(f(s,a))\Big\}, \\ \pi^{\shield}(s) &= \arg\max_{a\in\mathcal{A}} V(f(s,a)). \end{aligned} \tag{11}$$

To tractably approximate solutions to high-dimensional reachability problems, Fisac et al. [15] propose using reinforcement learning by replacing the standard Bellman equation with a time-discounted counterpart of 11:

$$V(s_t) = (1-\gamma)\ell(s) + \gamma \min\Big\{\ell_\theta(s), \max_{a\in\mathcal{A}} V\big(f(s,a)\big)\Big\}, \tag{12}$$

where γ is the discount factor that ensures contraction of the Bellman operator. The resulting *unsafe set*, denoted $\mathcal{U} \subset \mathcal{S}$, captures all states from which the system can no longer avoid entering $\mathcal{F}$, and is defined as the zero sublevel set of the value

function: $\mathcal{U} := \{s \mid V(s) < 0\}$. At deployment time, the safety value function and safety policy enable *safety filtering*: detecting unsafe actions proposed by any task policy π^{task} and minimally adjusting them only when necessary to ensure the system remains within the safe set.

C. Calibration Dataset

For each task, we collect a calibration dataset to determine the OOD threshold based on ensemble disagreement. This calibration dataset is a held-out subset collected alongside the training data, but it is not used during model training. Table III summarizes the calibration dataset sizes and the conformal prediction hyperparameters used for each task.

Task	Calibration Set Size (N)	α_{cal}	α_{trans}
Dubin's Car	500	0.05	0.05
Block Plucking	100	0.05	0.05
Jenga	30	0.10	0.10

TABLE III: Conformal Prediction Parameters for Each Task

D. Failure Detection of the Uncertainty-aware Safety Filter

Does our safety filter always guarantee safety? Our method assumes that the robot starts from an in-distribution initial state and maintains approximate control invariance with respect to an estimated safe set in the latent space. However, since the safety filter is trained via RL and relies on an imperfect latent dynamics model, safety cannot be guaranteed in all cases. The reliability of the learned filter can degrade in several situations—for instance, when the system begins in an out-of-distribution state (e.g., due to an OOD visual input at test time) or when the filter fails to prevent transitions into unsafe regions. In such cases, the safety filter may behave unpredictably, executing random or overconfident actions or even exacerbating unsafe situations. To ensure safe deployment, it is essential to detect when the safety filter becomes unreliable. In such cases, the system should halt and request human intervention. Without this safeguard, the robot may continue operating despite its internal safety mechanism failing.

System-level Failure Detection. Failures of learned safety filters can arise from a range of sources, including OOD sensory inputs, misspecified dynamics models, or inaccurately learned safety value functions. A reliable safety filter should exhibit consistent behavior under bounded epistemic uncertainty. To detect violations of this principle, we monitor whether the backup action $\pi^{\textcircled{\scriptsize V}}(z)$ leads to a transition with sufficiently low predictive uncertainty. If it does not, we assume the system has entered the OOD failure set and must stop operation.

Based on the safety filtering rule in Eq. 3, the selected action is expected to avoid transitions that induce high predictive uncertainty. Formally, the safety filter should satisfy: $\mathrm{D}\left(z, a^{\text{exec}}\right)) \leq \epsilon$. Conversely, if the filtered action itself leads to excessive epistemic uncertainty, we consider the system to have entered the unsafe set, which the robot cannot automatically recover from. In this case, the safety guarantees provided by the filter no longer hold, and the system should halt operation. In particular, if even the fallback action $\pi^{\textcircled{\scriptsize V}}(z)$ results in high disagreement, the system is deemed unrecoverable under the current safety filter: $\mathrm{D}(z, \pi^{\textcircled{\scriptsize V}}(z)) > \epsilon$. This motivates a modification to the filtering rule, introducing an explicit halting condition when the filter is unable to guarantee a safe and confident action. With predicted next latent state $\tilde{z}' \sim f_{\tilde{z}}(\tilde{z}, \pi^{\text{task}})$ the filter is constructed as:

$$\phi\left(\tilde{z}, \pi^{\text{task}}\right) := \begin{cases} \pi^{\text{task}}, & \text{if } V^{\textcircled{\scriptsize V}}(\tilde{z}')) > \delta, \\ \pi^{\textcircled{\scriptsize V}}(\tilde{z}), & \text{if } V^{\textcircled{\scriptsize V}}(\tilde{z}')) \leq \delta \;\&\; \mathrm{D}(z, \pi^{\textcircled{\scriptsize V}}(z)) \leq \epsilon, \\ \text{HALT}, & \text{otherwise.} \end{cases}$$

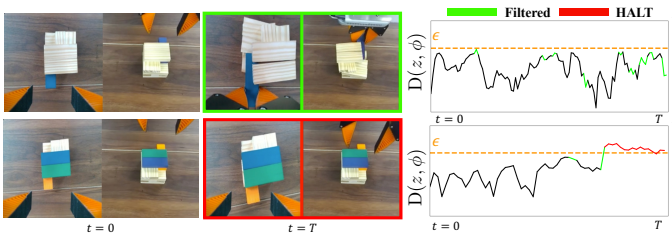

Fig. 7: *Top row*: despite a color change in the target block, the latent dynamics model remains reliable, maintaining predictive uncertainty below the threshold. *Bottom row*: in contrast, when the visual input deviates significantly from the training distribution, the model becomes unreliable. The safety filter fails to maintain predictive uncertainty below the threshold, prompting the system to halt in order to avoid actions that could compromise or aggravate safety.

Results: OOD Visual inputs. Fig. 7 illustrates the outcome of failure detection by the safety filter in the Jenga task. In this scenario, a teleoperator attempts to grasp a block and executes an unsafe action, pushing the block to the right. The learned safety filter intervenes to suppress this unsafe behavior. Although the block colors differ from those encountered during training, such visual changes do not inherently indicate OOD inputs. Instead, the decision to halt is governed by the reliability of the system. When the color of the target block changes but remains within the model's generalization capacity, the latent dynamics model remains reliable, maintaining predictive uncertainty below the threshold. In contrast, when the visual input deviates substantially from the training distribution, the model becomes unreliable. The safety filter then fails to keep uncertainty within acceptable bounds, prompting the system to halt in order to prevent potentially dangerous actions. Fig. 8 shows additional scenarios where the system safely halts upon detecting unrecoverable conditions due to OOD inputs that differ significantly from the training data.

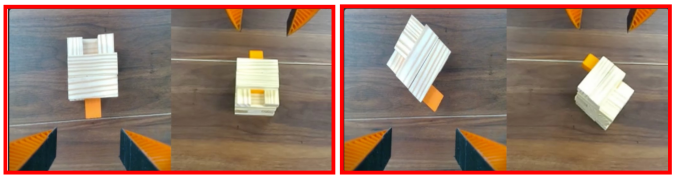

Fig. 8: OOD settings that lead the system to halt.