WISH: MARGIN-FREE LATENT CONTROL BARRIER FUNCTIONS VIA WORLD-MODEL IMAGINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Safety filters learned in world-model latent spaces commonly use binary failure labels to define unsafe states. However, their performance depends on the choice of a margin function that grades safety over nominally safe states, which is difficult to obtain in latent spaces. We propose WISH, a latent filter that measures the robustness of a latent state by the minimum cumulative disturbance effort required to reach a learned failure set against a fallback policy. Theoretically, for the exact modeled dynamics and value function, we prove that the resulting filter prevents failure when the initial state has sufficient disturbance resistance and the cumulative disturbance effort remains below a given budget. We also derive finite-horizon failure-probability bounds under stochastic imagination. In practice, WISH trains the latent safety value function entirely in world-model imagination using an adversary trained to minimize the disturbance effort required to cause failure against a fallback policy. Across four pixel-based continuous-control tasks, WISH achieves better safety-performance tradeoffs than existing latent filtering baselines.

1 INTRODUCTION

A safety filter must decide when to allow a task policy's action and when to intervene to avoid failure (Hsu et al., 2024). Recent work learns safety filters with Deep Reinforcement Learning (RL) in world-model latent spaces using failure classifiers to specify unsafe outcomes (Nakamura et al., 2025; 2026). We study how such filters can distinguish disturbance vulnerability and account for adverse outcomes when deciding whether to intervene. Recognizing failure, however, does not tell the filter how vulnerable a nominally safe state is. Two states may both be safe under a nominal policy, yet differ substantially in how robust they are to disturbances (Figure 1). To preserve safety while allowing task progress, the filter needs to distinguish states by their resistance to failure.

This presents two challenges. First, failure labels identify states to avoid but do not establish a meaningful ranking of safe states. Hamilton–Jacobi (HJ) reachability obtains a graded safety value by propagating a supplied continuous margin through the dynamics (Fisac et al., 2019), but correct failure classification does not ensure that the margin's magnitude reflects disturbance resistance. Learned latent coordinates also provide no prescribed physical metric from which to construct such a margin. Second, even a useful safety value can lead to unsafe decisions under stochastic dynamics. Expectation-based discrete-time barrier conditions support finite-horizon probabilistic safety bounds (Cosner et al., 2023), but can still admit substantial failure risk: favorable outcomes can compensate for failure outcomes in expectation. We therefore need both a meaningful robustness scale and a filtering condition that accounts for adverse outcomes.

Our key insight is that the least effort required to cause failure provides a measure of disturbance resistance and a basis for filtering actions according to adverse outcomes. Drawing on this, we propose **WISH** (**W**orld-model **I**magination for **S**afety without **H**and-crafted margins), a method for learning latent safety filters based on disturbance effort. Building on Least-Effort Adversarial Potentials (LEAP) (So et al., 2026), we develop a discrete-time formulation of the minimum cumulative disturbance effort required to reach a learned failure set against a fallback policy. For the exact value and a filter enforcing the corresponding admissibility condition over all disturbances, we prove avoidance of the modeled failure set whenever the initial value meets a prescribed budget and cumulative disturbance effort remains below it. Under the same conditions, we derive finite-horizon failure-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

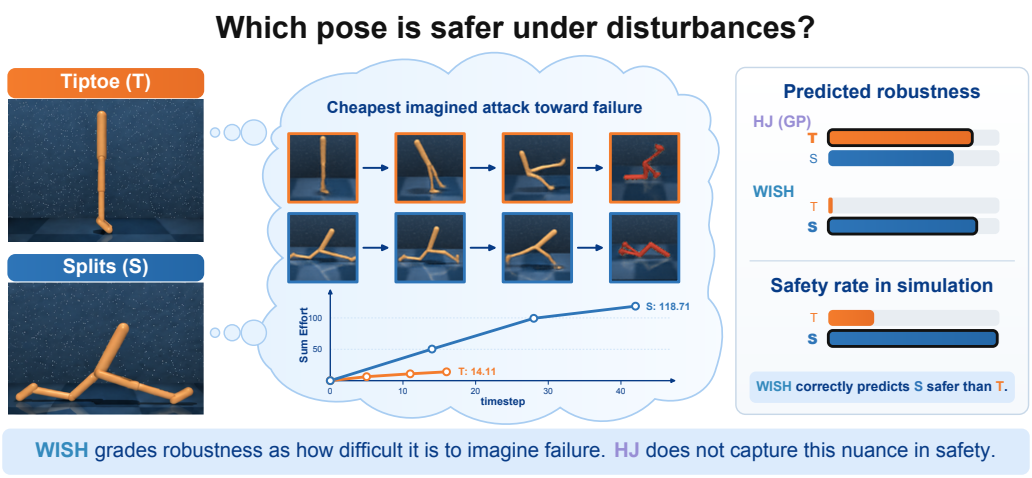

Figure 1: **WISH views safety as effort required to imagine failure.** Tiptoe and splits are both nominally safe walker poses, but a latent adversary requires substantially more cumulative effort to make the splits pose fail in imagination. HJ prefers tiptoe since the torso is physically further from the ground. However, when both poses are evaluated under noise, splits is the much more robust pose. WISH correctly identifies this difference by treating safety as how much effort is required to become unsafe in the imagined world model.

probability bounds for stochastic latent dynamics. WISH approximates this construction through adversarial learning in world-model imagination and sampled disturbance search; the theoretical guarantees do not certify these approximations. The deployed filtering threshold adjusts conservativeness without retraining the value function. Across four pixel-based continuous-control tasks, WISH improves the observed safety-performance tradeoffs over the evaluated baselines, providing empirical evidence for the utility of the learned value and sampled filter.

Contributions. We summarize our contributions as follows:

- We identify two challenges in latent safety filtering: accurate failure classification does not ensure a meaningful ranking of nominally safe states, and expectation-based safety constraints can admit actions with substantial failure risk even when the safety value is exact.
- These challenges inspire WISH, which measures safety by the minimum disturbance effort through the world model's noise channel required to enter the failure set and uses this value to construct a latent safety filter that accounts for adverse outcomes.
- We prove that the proposed filter theoretically certifies safety against disturbances whose cumulative effort remains below a user-specified budget, yielding finite-horizon failure-probability bounds under stochastic latent dynamics. The budget enables control over the filter's conservativeness without retraining the value function.
- We implement WISH with Deep RL using approximated values and demonstrate better empirical safety-performance tradeoffs than state-of-the-art baselines on latent control tasks defined with world models.

2 Related Work

We discuss the most relevant work here. See App. D for an extended discussion of related work.

Reachability and disturbance-based safety values. Safety filters connect safe control to reachability and barrier values (Hsu et al., 2024; Fisac et al., 2019; Choi et al., 2021). Least-Effort Adversarial Potentials (LEAP) measure the minimum cumulative disturbance effort required to defeat a fixed fallback and support continuous-time barrier-based filtering (So et al., 2026). We extend this construction to discrete-time latent dynamics, with filtered-policy probability bounds and learning in world-model imagination.

Safety in learned latent spaces. Latent Safety Filters extend HJ reachability to world-model representations with classifier-specified safety (Nakamura et al., 2025). LatentCBF addresses classifier-margin saturation and inaccurate nominal-action values through gradient penalties and broader training data (Nakamura et al., 2026). Both methods use HJ reachability to learn a value function that is used to construct a latent safety filter. WISH builds on these methods by identifying a more informative safety value and a corresponding filter that explicitly accounts for adverse outcomes, yielding improved safety-performance tradeoffs in our experiments.

Adversarial learning and recovery. Robust adversarial RL trains against destabilizing disturbances (Pinto et al., 2017), while ISAACS co-trains a fallback and disturbance policy for robust rollout certification (Hsu et al., 2023). Recovery RL already uses a safety critic to learn recovery and filter task actions (Thananjeyan et al., 2021). WISH learns a different quantity: the minimum cost of successful attacks through a frozen world model's noise channel.

Failure search and stochastic safety. Falsification and adaptive stress testing search for specification violations and likely failures (Corso et al., 2021; Lee et al., 2020). WISH uses the shifted negative log-likelihood as the attack cost for recovery and filtering. Its probability bounds follow from a cumulative-disturbance-budget implication, whereas stochastic barrier methods also provide bounds from expectation conditions under their assumptions (Cosner et al., 2023; 2024).

3 Latent Safety Filtering from Failure Labels

We consider the problem of latent-space safety filtering without a margin function.

3.1 Dynamics and Observations

We consider a discrete-time dynamical system with state $s_t \in \mathcal{S}$ controlled by actions $a_t \in \mathcal{A}$ and subject to disturbances. We assume that neither the state nor the disturbances can be observed directly. Instead, we have access to observations $o_t = o(s_t)$ with observation function $o : \mathcal{S} \to \mathcal{O}$. We use a world model (Ha & Schmidhuber, 2018b; Hafner et al., 2019) to maintain a latent representation $z_t \in \mathcal{Z} \subseteq \mathbb{R}^{n_z}$ from the observation-action history.

To evaluate the consequences of candidate actions, we need to predict how the system evolves without access to future observations. The world model provides an *imagination transition* g which predicts the next latent state from the current latent state and action. Using a reparameterized transition, we write the stochastic imagination dynamics as

$$z_{t+1} = g(z_t, a_t, \epsilon_t), \qquad \epsilon_t \sim q, \tag{1}$$

where $g : \mathcal{Z} \times \mathcal{A} \times \mathbb{R}^{n_\epsilon} \to \mathcal{Z}$ transforms noise from the base distribution q into latent successors. This formulation encapsulates a wide range of world models (Ha & Schmidhuber, 2018a; Hafner et al., 2019; 2020; 2023; Zhou et al., 2024).

Even when the physical dynamics are deterministic given the state, action, and disturbance, unobserved disturbances can make future observations uncertain. The stochastic imagination transition allows the world model to represent this uncertainty through the noise variable ϵ. Repeatedly applying g allows us to generate trajectories *in imagination* without access to future observations. We also consider a deterministic version of the imagination dynamics. Specifically, we fix the world model's reparameterization noise ϵ to zero and define

$$\bar{g}(z, a) \coloneqq g(z, a, 0). \tag{2}$$

Under this convention, deterministic imagination evolves as $z_{t+1} = \bar{g}(z_t, a_t)$.

3.2 Safety in Latent Space with Discrete Time CBFs

To evaluate safety in imagination, we assume access to a binary failure classifier $\hat{c} : \mathcal{Z} \to \{0, 1\}$. This classifier defines the latent failure set $\hat{\mathcal{F}} \coloneqq \{z \in \mathcal{Z} : \hat{c}(z) = 1\}$ to avoid in imagination.

In this work, we are interested in designing a safety filter (Hsu et al., 2024): given a nominal policy π_{nom}, construct a *filtered policy* π_{safe} that modifies the nominal policy only when necessary for

safety. While the classifier $\hat{c}$ identifies failure states, selecting safe actions requires accounting for how the system evolves. Discrete-time control barrier functions (D-CBFs) (Agrawal & Sreenath, 2017; Zeng et al., 2021) are a popular way to construct such a safety filter by enforcing a forward-invariance condition on a set of safe states. We first introduce this condition for the deterministic imagination dynamics $\bar{g}$.

Definition 1 (Discrete-time CBF (Agrawal & Sreenath, 2017; Zeng et al., 2021))**.** For a function $\hat{B} : \mathcal{Z} \to \mathbb{R}$, define the zero-superlevel set $\hat{\mathcal{C}} := \{z \in \mathcal{Z} : \hat{B}(z) \geq 0\}$. The function $\hat{B}$ is a D-CBF for $\bar{g}$, relative to $\hat{\mathcal{F}}$, if $\hat{\mathcal{C}} \cap \hat{\mathcal{F}} = \varnothing$ and there exists a class-$\mathcal{K}$ function $\alpha : [0, \infty) \to [0, \infty)$ with $\alpha(r) \leq r$ such that, for every $z \in \hat{\mathcal{C}}$, there exists an action $a \in \mathcal{A}$ satisfying

$$\hat{B}(\bar{g}(z, a)) - \hat{B}(z) \geq -\alpha(\hat{B}(z)). \tag{3}$$

Starting from $z_0 \in \hat{\mathcal{C}}$, selecting an action satisfying (3) at each step keeps the deterministic imagination trajectory within $\hat{\mathcal{C}}$ and outside the failure set $\hat{\mathcal{F}}$ (see App. B.1 for the standard argument). This lets us construct a safety filter by selecting an action close to the nominal action that satisfies (3). The D-CBF definition ensures that such an admissible action exists at every state in $\hat{\mathcal{C}}$.

Note that this guarantee concerns the deterministic imagination dynamics and does not by itself ensure safety of the physical system. Transferring the guarantee to physical execution requires additional consistency conditions relating the physical dynamics, the world model's latent updates, and the failure classifier (see App. A).

3.3 Challenge 1: The Missing Gradation with Hamilton-Jacobi Safety Filters

A popular approach to constructing D-CBFs is with HJ reachability (Mitchell et al., 2005; Zhang et al., 2025b), whose value function satisfies the deterministic backup

$$V_{\mathrm{HJ}}(z) = \min\Big\{h(z), \max_{a \in \mathcal{A}} V_{\mathrm{HJ}}(\bar{g}(z, a))\Big\}, \tag{4}$$

where the margin function[1] $h : \mathcal{Z} \to \mathbb{R}$ represents the latent failure set $\hat{\mathcal{F}} = \{z \in \mathcal{Z} : h(z) < 0\}$.

A key insight is that the margin h plays *two* roles: identifying failure states and grading safety outside the failure set. This gradation determines the worst margin along each trajectory, which HJ reachability optimizes. For filtering under disturbances, the ranking should reflect resistance to failure: a filter that treats fragile and robust states as equally safe has no basis for intervening when a nominal action reduces this resistance.

Constructing such a margin is particularly challenging when only binary failure labels are available. These labels identify failure states but do not specify how to rank states outside the failure set. Prior methods learn a margin using a hinge loss (Nakamura et al., 2025) and introduce gradient penalties to improve its smoothness (Nakamura et al., 2026). However, correct classification and smoothness alone do not ensure that the resulting margin captures disturbance resistance. We illustrate this distinction in Fig. 1. Both poses permit nominally safe operation, but the splits state is more robust under noise than the tiptoe state. The learned HJ values are either indifferent to the two poses, or prefer the less robust tiptoe state, showing that a margin learned from failure labels need not produce a value aligned with disturbance resistance.

3.4 Challenge 2: Safety under Stochastic Imagination

Beyond constructing a meaningful safety value, we must determine which actions preserve safety under stochastic transitions. The deterministic D-CBF condition constrains a single successor under $\bar{g}$, whereas stochastic imagination produces a distribution of possible successors. Thus, satisfying the deterministic condition alone does not guarantee safety under stochastic imagination. Prior work on latent safety filters adopts this approach by replacing the deterministic HJ backup with a stochastic backup (Nakamura et al., 2026).

$$V_{\mathrm{HJ}}(z) = \min\Big\{h(z), \max_{a \in \mathcal{A}} \mathbb{E}[V_{\mathrm{HJ}}(g(z, a, \epsilon))]\Big\}, \tag{5}$$

[1] also called an implicit surface function (Mitchell et al., 2005) or a target function (Bansal & Tomlin, 2021)

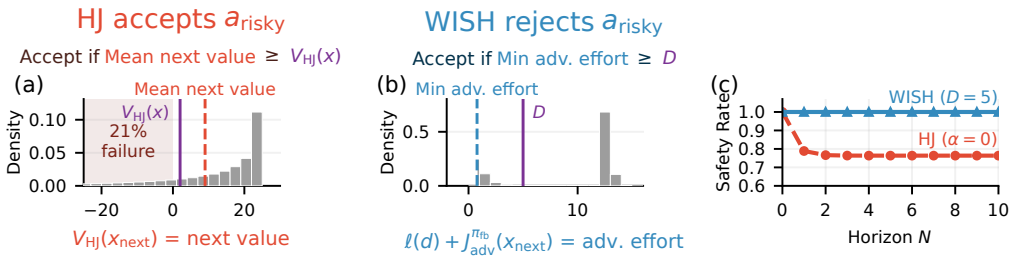

Figure 2: **Even the strictest HJ mean constraint allows high failure risk.** We compare the strictest ($\alpha = 0$) HJ filter with WISH ($D = 5$) on a toy example (App. B.11). Both filter the same π_{nom} and can apply a safe action with near-zero failure rate. **(a)** HJ accepts a nominal action with a 21% one-step failure rate because its expected successor value satisfies the mean constraint. **(b)** WISH rejects this action because its minimum adversarial effort falls below the required budget, selecting the safer fallback and **(c)** achieving near-perfect safety rate.

The corresponding filter constrains the expected decrease in the safety value, yielding finite-horizon safety bounds under the assumptions in App. B.10. However, this condition permits compensation between outcomes: large values on favorable successors can offset negative values on failure successors. This remains possible even when the filter permits no decrease in expectation:

$$\mathbb{E}_{\epsilon}[V_{\text{HJ}}(g(z, a, \epsilon))] - V_{\text{HJ}}(z) \geq 0, \tag{6}$$

Fig. 2 illustrates this limitation. The risky action satisfies the expected-nondecrease condition (6) despite having a 21% one-step failure probability because favorable outcomes compensate for the failure outcomes. The filter therefore accepts this action when proposed by the nominal policy, even though an alternative action has near-zero failure probability.

4 Uncertainty-Aware Safety Filtering with WISH

4.1 Measuring Safety as Robustness to Disturbances

To distinguish states by their robustness, we measure how much disturbance is required to cause failure: a state is safer if reaching failure requires greater disturbance effort. Since this effort depends on how the controller responds, we fix a fallback policy $\pi_{\text{fb}} : \mathcal{Z} \to \mathcal{A}$ and measure the minimum cumulative disturbance effort required to drive the system into failure under this policy. This follows the Least-Effort Adversarial Potentials (LEAP) construction (So et al., 2026), which we extend to discrete-time latent dynamics.

To measure resistance to world-model stochasticity, we let an adversary *choose* realizations of the imagination noise ϵ. We measure disturbance effort by its negative log-likelihood, shifted so that zero noise has zero cost:

$$\xi(\epsilon) := -\log q(\epsilon) + \log q(0). \tag{7}$$

Assuming q is positive and finite with a mode at zero, this cost is nonnegative and assigns greater effort to less likely noise realizations. For standard Gaussian noise, this gives $\xi(\epsilon) = \frac{1}{2}\|\epsilon\|^2$.

Let τ denote the first hitting time of the failure set $\hat{\mathcal{F}}$, i.e., $\tau := \inf\{t \geq 0 : z_t \in \hat{\mathcal{F}}\}$. We now define the minimum disturbance cost to failure $J^{\pi_{\text{fb}}}_{\text{adv}}(z)$ for a state $z \in \mathcal{Z}$ as the minimum cumulative disturbance cost required to drive the system into failure under the fallback policy π_{fb}, which can be written as the following *deterministic* least-effort adversarial optimal control problem

$$J^{\pi_{\text{fb}}}_{\text{adv}}(z_0) := \inf_{\epsilon_{0:\infty}} \left\{ \sum_{t=0}^{\tau-1} \xi(\epsilon_t) : \tau < \infty \right\}, \quad z_{t+1} = g(z_t, \pi_{\text{fb}}(z_t), \epsilon_t). \tag{8}$$

We take $\inf \emptyset = +\infty$, so $J^{\pi_{\text{fb}}}_{\text{adv}}(z_0) = +\infty$ if no disturbance sequence can reach failure from z_0. For conciseness, we assume throughout the main-text analysis that $J^{\pi_{\text{fb}}}_{\text{adv}}$ is finite everywhere; App. B.9 explains how to handle extended-valued $J^{\pi_{\text{fb}}}_{\text{adv}}$. $J^{\pi_{\text{fb}}}_{\text{adv}}$ is zero on $\hat{\mathcal{F}}$, as well as any state that reaches failure under π_{fb} without any disturbance. Moreover, note that (8) only requires a binary failure classifier $\hat{c}$ to determine whether a state is in the failure set $\hat{\mathcal{F}}$, and does not require any additional information about the safety of a state, unlike HJ reachability.

Safety properties of $J^{\pi_{\text{fb}}}_{\text{adv}}$. A key property of $J^{\pi_{\text{fb}}}_{\text{adv}}$ is that it bounds the decrease of the value under the fallback policy π_{fb} and any disturbance $\epsilon \in \mathbb{R}^{n_\epsilon}$.

$$J^{\pi_{\text{fb}}}_{\text{adv}}(g(z, \pi_{\text{fb}}(z), \epsilon)) \geq J^{\pi_{\text{fb}}}_{\text{adv}}(z) - \xi(\epsilon), \quad \forall z \in \mathcal{Z}, \epsilon \in \mathbb{R}^{n_\epsilon}. \tag{9}$$

This allows $J^{\pi_{\text{fb}}}_{\text{adv}}$ to satisfy several desirable safety properties. Let $\alpha(r)$ be a class-$\mathcal{K}$ function with $\alpha(r) \leq r$ for all $r \geq 0$.

Lemma 1. *For $D > 0$, define $V_D \coloneqq J^{\pi_{\text{fb}}}_{\text{adv}} - D$ as a shifted version of the value function, and let $\mathcal{S}_D \coloneqq \{z \in \mathcal{Z} : J^{\pi_{\text{fb}}}_{\text{adv}}(z) \geq D\}$ denote the D-superlevel set of $J^{\pi_{\text{fb}}}_{\text{adv}}$. Then, V_D is a D-CBF (Def. 1) for the deterministic latent dynamics* (2)*, i.e.,*

$$V_D(g(z, \pi_{\text{fb}}(z), 0)) - V_D(z) \geq 0 \geq -\alpha(V_D(z)), \quad \forall z \in \mathcal{S}_D. \tag{10}$$

We put all proofs in App. B. Write $c(\epsilon_{0:t}) = \sum_{k=0}^{t} \xi(\epsilon_k)$ for the cumulative disturbance cost. Next, $J^{\pi_{\text{fb}}}_{\text{adv}}$ quantifies how robust a state is to disturbances in terms of this cost.

Lemma 2. *For any state $z \in \mathcal{S}_D$, $J^{\pi_{\text{fb}}}_{\text{adv}}$ certifies that any disturbance sequence $\epsilon_{0:t}$ with cumulative cost $c(\epsilon_{0:t}) < D$ does not drive the system into failure under the fallback policy π_{fb}.*

4.2 Constructing a Safety Filter

Lem. 2 assumes that the fallback policy acts at every step. A safety filter must instead determine whether a proposed nominal action preserves sufficient resistance to failure. To evaluate a candidate action a, we consider the minimum disturbance cost required to cause failure when taking a once and following the fallback thereafter by defining the action-value function Q.

$$Q(z, a) \coloneqq \inf_{\epsilon \in \mathbb{R}^{n_\epsilon}} \left\{ \xi(\epsilon) + J^{\pi_{\text{fb}}}_{\text{adv}}(g(z, a, \epsilon)) \right\} \tag{11}$$

Let $Q_D \coloneqq Q - D$ denote the shifted Q function. We define the admissible action set $\mathcal{A}_{V_D}$ as

$$\mathcal{A}_{V_D}(z) \coloneqq \{a \in \mathcal{A} : Q_D(z, a) - V_D(z) \geq -\alpha(\max\{V_D(z), 0\})\}. \tag{12}$$

When $J^{\pi_{\text{fb}}}_{\text{adv}}(z) \geq D$, (12) allows $J^{\pi_{\text{fb}}}_{\text{adv}}$ to decrease, controlled by rate α, while ensuring $Q(z, a) \geq D$. Otherwise, when $J^{\pi_{\text{fb}}}_{\text{adv}}(z) < D$, it requires $Q(z, a) \geq J^{\pi_{\text{fb}}}_{\text{adv}}(z)$, accepting only actions that preserve at least as much disturbance robustness as the fallback action $\pi_{\text{fb}}(z)$.

The following result extends Lem. 2 to policies selecting admissible actions.

Lemma 3. *Let $z_0 \in \mathcal{S}_D$ and let π satisfy $\pi(z) \in \mathcal{A}_{V_D}(z)$ for every $z \in \mathcal{Z}$, with $\mathcal{A}_{V_D}$ defined in* (12)*. Then, for every disturbance sequence $\epsilon_{0:t}$ for $t \geq 0$, $c(\epsilon_{0:t}) < D \implies z_{t+1} \notin \hat{\mathcal{F}}$. Moreover, $\pi_{\text{fb}}(z) \in \mathcal{A}_{V_D}(z)$ for every $z \in \mathcal{Z}$.*

Since the fallback supplies an admissible action, we construct a safety filter π_{V_D} by projecting the nominal action onto $\mathcal{A}_{V_D}$, assuming a minimizer exists:

$$\pi_{V_D}(z, \pi_{\text{nom}}) \coloneqq \operatorname*{arg\,min}_{a \in \mathcal{A}_{V_D}(z)} \|a - \pi_{\text{nom}}(z)\|. \tag{13}$$

Under stochastic imagination dynamics (1), the noise ϵ_t is sampled from q rather than chosen by the adversary. The disturbance-budget guarantee allows us to bound the probability of failure in terms of the probability that the disturbances exceed the budget D.

Theorem 1. *Let $z_0 \in \mathcal{S}_D$ and $\pi(z) \in \mathcal{A}_{V_D}(z),\ \forall z \in \mathcal{Z}$. Then, $\mathbb{P}(\tau \leq N) \leq \mathbb{P}(c(\epsilon_{0:N-1}) \geq D)$.*

For independent standard-normal imagination noise and quadratic disturbance cost, the cumulative cost has a scaled χ^2 distribution. Applying Thm. 1 to the WISH filter π_{V_D} therefore gives the following explicit bound; see App. B.8 for the derivation.

Corollary 1. *Suppose the imagination noises are independent across time with standard normal distribution $q = \mathcal{N}(0, I_{n_\epsilon})$, and $\xi(\epsilon) = \frac{1}{2}\|\epsilon\|^2$. Then, for any initial state $z_0 \in \mathcal{S}_D$, under the safety filter π_{V_D} and the stochastic imagination dynamics* (1)*, the probability of failing within N steps is bounded by*

$$\mathbb{P}\Big(\tau \leq N\Big) \leq 1 - F_{\chi^2}\left(2D; N n_\epsilon\right), \qquad 0 < D \leq J^{\pi_{\text{fb}}}_{\text{adv}}(z_0), \tag{14}$$

where $F_{\chi^2}(\cdot; k)$ is the CDF of the χ^2 distribution with k degrees of freedom, and n_ϵ is the dimension of the disturbance ϵ.

Controllable safety with the WISH filter. The user-specified budget $D > 0$ sets the disturbance resistance of the WISH filter (13). Increasing D reduces the failure-probability upper bound in Thm. 1 up to the maximum value $D = J_{\mathrm{adv}}^{\pi_{\mathrm{fb}}}(z_0)$. Returning to the toy example in Fig. 2, a sufficiently large budget rejects the risky nominal action and selects the safer fallback, achieving near-perfect safety where the HJ expectation-based filter continues to accept the risky action.

4.3 Learning a WISH filter with Reinforcement Learning in Imagination

Having defined the WISH filter, we now propose a method of learning the value function $J_{\mathrm{adv}}^{\pi_{\mathrm{fb}}}$, which requires solving the least-effort adversarial optimal control problem (8). Unlike the HJ optimal control problem (Nakamura et al., 2026), which is solved using an HJ fixed-point equation (Fisac et al., 2019), (8) has a sum-over-time objective, allowing us to use standard RL algorithms to learn a policy for the adversary. The only challenge is the constraint that the adversary must drive the system into the failure set $\hat{\mathcal{F}}$. We use a discounted disturbance-cost objective as a heuristic surrogate for this constrained problem. Drawing inspiration from constrained optimization techniques, we learn a classifier that predicts whether the adversary can reach $\hat{\mathcal{F}}$ and penalize the adversary for failing to reach $\hat{\mathcal{F}}$ (see App. C.3). As with prior work on learning neural safety filters with HJ reachability (So et al., 2024b; Oh et al., 2025; Nakamura et al., 2026), discounting facilitates stable value learning. We implement the adversary with Proximal Policy Optimization (PPO) (Schulman et al., 2017): its critic V_{adv} estimates the discounted return under the per-step reward $-\xi(\epsilon_t)$, and we use $\widehat{J} = -V_{\mathrm{adv}}$ as an empirical proxy for disturbance resistance. This objective alone does not enforce eventual failure: discounting can favor delayed attacks. However, our experiments support the utility of this proxy for safety filtering and fallback learning. A more principled treatment of the failure-reaching constraint, for example using the methods of Liao & Liang (2021); So et al. (2024a), remains future work.

Impact of the Feedback Policy π_{fb}. Although the theoretical results in Sec. 4.1 hold for any feedback policy π_{fb}, its quality significantly affects the resulting WISH filter (13). A weak fallback policy that fails under zero disturbance $\epsilon = 0$ yields $J_{\mathrm{adv}}^{\pi_{\mathrm{fb}}} = 0$ on those initial states, making the filter overly conservative. When a strong fallback policy is unavailable, this motivates learning one that is robust to disturbances. As discussed earlier, HJ reachability is challenging when the margin function h poorly captures safety, especially when only a binary failure classifier $\hat{c}$ is available. The value function $J_{\mathrm{adv}}^{\pi_{\mathrm{fb}}}$, however, measures robustness to disturbances using only $\hat{c}$. We therefore propose learning π_{fb} with HJ reachability using the learned $J_{\mathrm{adv}}^{\pi_{\mathrm{fb}}}$ as the margin function h.

One issue is that HJ reachability requires $z \in \hat{\mathcal{F}} \implies h(z) < 0$, whereas $J_{\mathrm{adv}}^{\pi_{\mathrm{fb}}}$ is zero on $\hat{\mathcal{F}}$. In fact, the learned approximation to $J_{\mathrm{adv}}^{\pi_{\mathrm{fb}}}$ may have errors on $\hat{\mathcal{F}}$ that make it strictly positive on some failure states. If the estimated values on $\hat{\mathcal{F}}$ are bounded above, subtracting a sufficiently large constant makes them negative there. For exact HJ value iteration, this shift changes the values but preserves the maximizing actions, so it is unnecessary for selecting the fallback policy.

Lemma 4. *Adding a constant c to the margin h shifts the HJ value function in* (4) *or* (5) *by c without changing the maximizing actions.*

We now present WISH (World-model Imagination for Safety without Hand-crafted margins), an approach to learning safety filters under stochastic latent dynamics in imagination (see Alg. 1). We start with a fallback policy $\pi_{\mathrm{fb},0}$ (we use the zero policy) and solve the least-effort adversarial problem under $\pi_{\mathrm{fb},0}$ to obtain $J_{\mathrm{adv}}^{\pi_{\mathrm{fb},0}}$. Then, we use $J_{\mathrm{adv}}^{\pi_{\mathrm{fb},0}}$ as the margin function h to learn a new fallback policy π_{fb} by solving an HJ reachability problem (4). We then solve the least-effort adversarial problem but under π_{fb} to obtain $J_{\mathrm{adv}}^{\pi_{\mathrm{fb}}}$. Finally, we construct the WISH filter π_{V_D} (13) using the learned $J_{\mathrm{adv}}^{\pi_{\mathrm{fb}}}$. At deployment, we approximate the $\inf_\epsilon$ in Q (11) with samples and use the resulting estimates to test candidate actions for admissibility (see App. C.3 for details). If no candidate satisfies the estimated admissibility condition, we execute the fallback policy π_{fb}.

5 Experiments

We conduct experiments to answer the following research questions. **(Q1)** How does WISH compare with existing latent-space safety filters in balancing safety and task performance? **(Q2)** Does mea-

Algorithm 1 WISH training in world-model imagination

Require: frozen world model $(g, \hat{c})$ and latent training-state sampler

1: Initialize fallback $\pi_{\text{fb},0}$ and adversary $(\pi_{\text{adv}}, V_{\text{adv}})$
2: Train $(\pi_{\text{adv}}, V_{\text{adv}})$ against frozen $\pi_{\text{fb},0}$
3: Train π_{fb} against frozen $\widehat{J} = -V_{\text{adv}}$
4: Train $(\pi_{\text{adv}}, V_{\text{adv}})$ against frozen π_{fb}
5: **return** π_{fb} and $\widehat{J} = -V_{\text{adv}}$

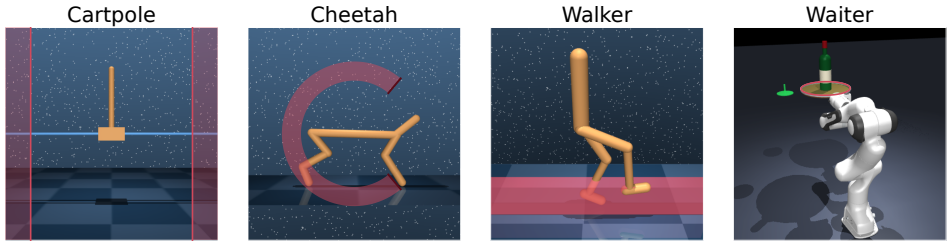

Figure 3: **Evaluation tasks and their failure sets.** Four world-model tasks (Cartpole, Cheetah, Walker, Waiter). Failure sets are illustrated in **red**.

suring safety through disturbance effort enable better safety-performance tradeoffs than expectation-based HJ reachability independent of fallback policy quality? **(Q3)** Does learning a fallback policy using disturbance-based safety values produce more robust recovery behavior than conventional HJ fallback learning? **(Q4)** Does WISH provide a more informative gradation of safety than classifier-derived margin functions? We provide implementation details, task descriptions, and additional results in App. C.

Baselines. We compare WISH against a Nominal task policy, two learned Hamilton-Jacobi reachability filters in latent space labeled HJ (GP) (Nakamura et al., 2026) and HJ (no GP) (Nakamura et al., 2025), and a model-predictive shielding (MPS) filter (Bastani, 2021). For MPS, we default the fallback control to that of HJ (GP). Details for each baseline are given in App. C.2.

Benchmarks. As illustrated in Fig. 3, we evaluate on four world-model tasks in MuJoCo (Todorov et al., 2012): Cartpole, Cheetah, Walker, and Waiter. Each task defines a performance and safety component (App. C.1) which induces a tradeoff between both objectives.

Evaluation Metrics. Since filter parameters (e.g., robustness threshold D and D-CBF α) inherently induce a tradeoff between safety and performance, we sweep each baseline's hyperparameters to produce a safety-performance Pareto frontier. For each frontier plot, we report the safety rate (fraction of episodes that don't enter the failure set) and each task's performance metric.

5.1 Results

(Q1) WISH improves safety-performance tradeoffs across all four tasks. We present the safety-performance Pareto frontiers for all tasks in Fig. 4. Across all tasks, WISH achieves higher task performance at comparable safety levels than the baselines.

(Q2) WISH's safety gains extend beyond improvements in the fallback policy. On Cheetah, WISH achieves higher safety than HJ (GP) when both use the HJ (GP) fallback policy (Fig. 5a). Because the fallback is shared, this supports disturbance-effort value learning and filtering improving upon HJ (GP) beyond yielding a better fallback. Full WISH further improves the safety-performance tradeoff, suggesting that its learned fallback does still provide benefits.

(Q3) WISH values provide more useful guidance for learning robust fallback policies. On Cheetah, the fallback learned using the WISH value achieves higher safety under disturbances than the fallback learned using the classifier-derived margin (Fig. 5b). This suggests that the disturbance-based value provides a more informative learning signal for selecting robust fallback actions.

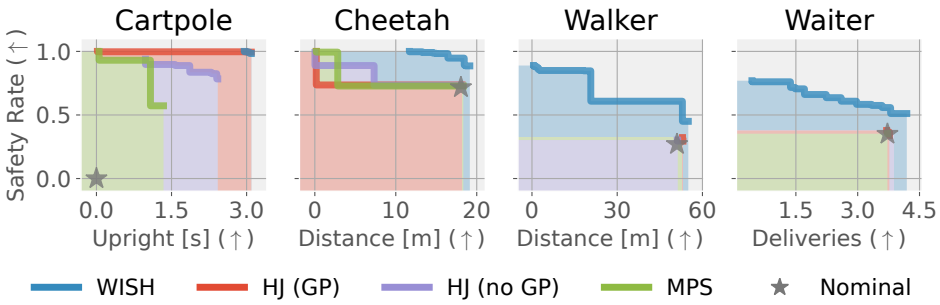

Figure 4: **Safety-performance Pareto frontiers.** We evaluate WISH and baselines on the tasks in Fig. 3 comparing task performance and safety. Each point represents a different choice of filter hyperparameters. WISH consistently dominates the baselines across all tasks, achieving higher performance at the same safety level or higher safety at the same performance level.

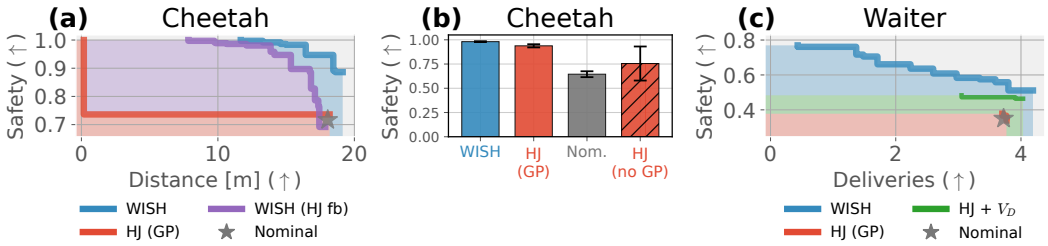

Figure 5: **Ablations. (a)** Safety-performance Pareto frontiers on Cheetah with a WISH variant that trains with the HJ (GP) fallback. **(b)** Average safety rates of each baseline's learned fallback policies evaluated over sampled Cheetah environment states. **(c)** Safety-performance Pareto frontiers on Waiter with an HJ (GP) variant that uses the WISH adversary value V_D as its margin function.

(Q4) The learned safety margin also improves HJ filtering. Beyond guiding fallback learning, we test whether the disturbance-based value also improves the safety-performance tradeoff when used as the margin for an HJ filter. On Waiter, replacing the classifier-derived margin with the WISH value improves the observed frontier (Fig. 5c). This demonstrates that the value's utility extends to expectation-based HJ filtering.

6 Conclusion and Limitations

We introduced WISH, which measures safety by the minimum disturbance effort required to enter the failure set under only binary failure information and world-model dynamics. We identified two challenges in latent safety filtering: failure classification alone does not establish a meaningful safety gradation, and expectation-based filtering can admit substantial failure risk. WISH addresses both using LEAP's minimum disturbance effort (So et al., 2026) to grade safety and construct a filter that preserves disturbance resistance. We prove the resulting filter guarantees safety below a cumulative disturbance budget, yielding finite-horizon failure-probability bounds under stochastic imagination. Experiments show improved safety-performance tradeoffs over the evaluated baselines and support disturbance effort as a useful signal for safety gradation and recovery learning.

Limitations. The exact guarantees do not hold for the learned approximations without additional verification (Zhang et al., 2023). The discounted adversary objective does not explicitly enforce successful failure, so even exact optimization of this surrogate need not correspond to the solution in (8). The learned adversary and fallback may also go out of distribution and underperform when the world model is inaccurate. For example, the adversary may exploit model errors to imagine unrealistic attacks, producing an uninformative safety value. Finally, our probability bounds concern stochastic imagination under the modeled latent transition law and do not directly apply to noisy physical execution with observation-conditioned posterior updates. We leave this for future work.

AI DISCLOSURE

In this work, we used generative AI tools for help with code implementation, brainstorming, and paper polishing and writing assistance. We also used generative AI tools to help document our algorithmic approach and hyperparameters, including translating implementation details into consistent notation and equations. We checked these descriptions against the code and experimental configuration. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Complete proofs of the theoretical results are provided in App. B. Task definitions and failure sets are detailed in App. C.1, and the evaluated baselines are described in App. C.2. Model architectures, training schedules and hyperparameters, classifier fitting, policy evaluation, deployment-time filtering, and the evaluation protocol are specified in App. C.3 and Tab. 1. All methods use the same frozen world models and nominal policies and are evaluated with matched initial conditions and disturbance sequences where applicable.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning*, pp. 22–31, 2017. URL https://proceedings.mlr.press/v70/achiam17a.html.

Ayush Agrawal and Koushil Sreenath. Discrete control barrier functions for safety-critical control of discrete systems with application to bipedal robot navigation. In *Robotics: Science and Systems (RSS)*, 2017.

Sankalp Agrawal, Junwon Seo, Kensuke Nakamura, Ran Tian, and Andrea Bajcsy. AnySafe: Adapting latent safety filters at runtime via safety constraint parameterization in the latent space. In *2026 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2026. URL https://any-safe.github.io/.

Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In *AAAI Conference on Artificial Intelligence*, pp. 2669–2678, 2018. URL https://ojs.aaai.org/index.php/AAAI/article/view/11797.

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.

Aaron D. Ames, Samuel Coogan, Magnus Egerstedt, Gennaro Notomista, Koushil Sreenath, and Paulo Tabuada. Control barrier functions: Theory and applications. In *2019 18th European Control Conference (ECC)*, pp. 3420–3431, 2019.

Yashwanth Annapureddy, Che Liu, Georgios Fainekos, and Sriram Sankaranarayanan. S-TaLiRo: A tool for temporal logic falsification for hybrid systems. In *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, pp. 254–257. Springer, 2011. URL https://home.cs.colorado.edu/~srirams/papers/sTaliro-tacas11.pdf.

Yarden As, Ilnura Usmanova, Sebastian Curi, and Andreas Krause. Constrained policy optimization via bayesian world models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=PRZoSmCinhf.

Jean-Pierre Aubin, Alexandre M. Bayen, and Patrick Saint-Pierre. *Viability Theory: New Directions*. Springer, 2011. URL https://doi.org/10.1007/978-3-642-16684-6.

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

Somil Bansal, Mo Chen, Sylvia Herbert, and Claire J. Tomlin. Hamilton-jacobi reachability: A brief overview and recent advances. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pp. 2242–2253, 2017.

Osbert Bastani. Safe reinforcement learning with nonlinear dynamics via model predictive shielding. In *2021 American control conference (ACC)*, pp. 3488–3494. IEEE, 2021.

Osbert Bastani, Shuo Li, and Anton Xu. Safe reinforcement learning via statistical model predictive shielding. In *Robotics: Science and Systems*, volume 17, pp. 10–15607, 2021.

Andrew Bennett, Dipendra Misra, and Nathan Kallus. Provable safe reinforcement learning with binary feedback. In *International Conference on Artificial Intelligence and Statistics*, pp. 10871–10900, 2023. URL https://proceedings.mlr.press/v206/bennett23a.html.

Lukas Brunke, Melissa Greeff, Adam W. Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P. Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 5:411–444, 2022. URL https://arxiv.org/abs/2108.06266.

Steven Carr, Nils Jansen, Sebastian Junges, and Ufuk Topcu. Safe reinforcement learning via shielding under partial observability. In *AAAI Conference on Artificial Intelligence*, pp. 14748–14756, 2023. URL https://ojs.aaai.org/index.php/AAAI/article/view/26723.

Yuxiao Chen, Mrdjan Jankovic, Mario Santillo, and Aaron D. Ames. Backup control barrier functions: Formulation and comparative study. In *IEEE Conference on Decision and Control*, 2021. URL https://arxiv.org/abs/2104.11332.

Richard Cheng, Gábor Orosz, Richard M. Murray, and Joel W. Burdick. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In *AAAI Conference on Artificial Intelligence*, pp. 3387–3395, 2019. URL https://ojs.aaai.org/index.php/AAAI/article/view/4213.

Jason J Choi, Donggun Lee, Koushil Sreenath, Claire J Tomlin, and Sylvia L Herbert. Robust control barrier-value functions for safety-critical control. In *IEEE Conference on Decision and Control (CDC)*, pp. 6814–6821. IEEE, 2021.

Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, 18(167): 1–51, 2018.

Anthony Corso, Robert J Moss, Mark Koren, Ritchie Lee, and Mykel J Kochenderfer. A survey of algorithms for black-box safety validation of cyber-physical systems. *Journal of Artificial Intelligence Research*, 72:377–428, 2021. URL https://arxiv.org/abs/2005.02979.

Ryan K Cosner, Preston Culbertson, Andrew J Taylor, and Aaron D Ames. Robust safety under stochastic uncertainty with discrete-time control barrier functions. *arXiv preprint arXiv:2302.07469*, 2023.

Ryan K. Cosner, Preston Culbertson, and Aaron D. Ames. Bounding stochastic safety: Leveraging Freedman's inequality with discrete-time control barrier functions. *IEEE Control Systems Letters*, 2024. URL https://arxiv.org/abs/2403.05745.

Gal Dalal, Krishnamurthy Dvijotham, Matej Vecerik, Todd Hester, Cosmin Paduraru, and Yuval Tassa. Safe exploration in continuous action spaces. *arXiv preprint arXiv:1801.08757*, 2018. URL https://arxiv.org/abs/1801.08757.

Charles Dawson, Sicun Gao, and Chuchu Fan. Safe control with learned certificates: A survey of neural lyapunov, barrier, and contraction methods for robotics and control. *IEEE Transactions on Robotics*, 39(3):1749–1767, 2023. URL https://arxiv.org/abs/2202.11762.

Sarah Dean, Andrew J. Taylor, Ryan K. Cosner, Benjamin Recht, and Aaron D. Ames. Guaranteeing safety of learned perception modules via measurement-robust control barrier functions. In *Proceedings of the 2020 Conference on Robot Learning*, pp. 654–670, 2021. URL https://proceedings.mlr.press/v155/dean21a.html.

Alexandre Donzé. Breach, a toolbox for verification and parameter synthesis of hybrid systems. In *Computer Aided Verification*, pp. 167–170, 2010. URL https://doi.org/10.1007/978-3-642-14295-6_17.

Weinan E, Weiqing Ren, and Eric Vanden-Eijnden. Minimum action method for the study of rare events. *Communications on Pure and Applied Mathematics*, 57(5):637–656, 2004. URL https://doi.org/10.1002/cpa.20005.

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

Milan Ganai, Sicun Gao, and Sylvia L Herbert. Hamilton-jacobi reachability in reinforcement learning: A survey. *IEEE Open Journal of Control Systems*, 3:310–324, 2024.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018a.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3):440, 2018b.

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, 2019.

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.

Kai-Chieh Hsu, Duy Phuong Nguyen, and Jaime Fernández Fisac. ISAACS: Iterative soft adversarial actor-critic for safety. In *Learning for Dynamics and Control Conference*, pp. 90–103. PMLR, 2023.

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*, 7:47–72, 2024.

Garud N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005. URL https://doi.org/10.1287/moor.1040.0129.

Matthew Kim, Kensuke Nakamura, and Andrea Bajcsy. How well do latent world models understand partially observable safety constraints? *arXiv preprint arXiv:2510.06492*, 2026. URL https://arxiv.org/abs/2510.06492.

Shishir Kolathaya and Aaron D. Ames. Input-to-state safety with control barrier functions. *IEEE Control Systems Letters*, 3(1):108–113, 2019. URL https://arxiv.org/abs/1803.03035.

Mark Koren, Ahmed Nassar, and Mykel J Kochenderfer. Finding failures in high-fidelity simulation using adaptive stress testing and the backward algorithm. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021. URL https://arxiv.org/abs/2107.12940.

Ritchie Lee, Ole J Mengshoel, Anshu Saksena, Ryan Gardner, Daniel Genin, Joshua Silbermann, Michael Owen, and Mykel J Kochenderfer. Adaptive stress testing: Finding likely failure events with reinforcement learning. *Journal of Artificial Intelligence Research*, 69:1165–1201, 2020. URL https://arxiv.org/abs/1811.02188.

Jingqi Li, Donggun Lee, Somayeh Sojoudi, and Claire J Tomlin. Infinite-horizon reach-avoid zero-sum games via deep reinforcement learning. *arXiv preprint arXiv:2203.10142*, 2022.

Zeyang Li, Chuxiong Hu, Shengbo Eben Li, Jia Cheng, and Yunan Wang. Robust safe reinforcement learning under adversarial disturbances. *arXiv preprint arXiv:2310.07207*, 2023. URL https://arxiv.org/abs/2310.07207.

Weiwei Liao and Tao Liang. Computation of reachable sets based on hamilton-jacobi-bellman equation with running cost function. *arXiv preprint arXiv:2107.11941*, 2021.

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. URL https://arxiv.org/abs/2111.09971.

Jingyu Liu, Yuanda Wang, Changyin Sun, and Anqing Duan. Safe online reinforcement learning with diffusion world model and langevin dynamics. *Expert Systems with Applications*, 310: 131278, 2026. URL https://doi.org/10.1016/j.eswa.2026.131278.

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.

Kensuke Nakamura, Lasse Peters, and Andrea Bajcsy. Generalizing safety beyond collision-avoidance via latent-space reachability analysis. In *Robotics: Science and Systems*, 2025. doi: 10.15607/RSS.2025.XXI.113.

Kensuke Nakamura, Arun L Bishop, Steven Man, Aaron M Johnson, Zachary Manchester, and Andrea Bajcsy. How to train your latent control barrier function: Smooth safety filtering under hard-to-model constraints. In *Learning for Dynamics and Control Conference*, pp. 145–168. PMLR, 2026.

Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 2005. URL https://people.eecs.berkeley.edu/~elghaoui/Pubs/RobMDP_OR2005.pdf.

Donggeon David Oh, Justin Lidard, Haimin Hu, Himani Sinhmar, Elle Lazarski, Deepak Gopinath, Emily S Sumner, Jonathan A DeCastro, Guy Rosman, Naomi Ehrich Leonard, et al. Safety with agency: Human-centered safety filter with application to ai-assisted motorsports. *arXiv preprint arXiv:2504.11717*, 2025.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *International Conference on Machine Learning*, pp. 2817–2826. PMLR, 2017. URL https://proceedings.mlr.press/v70/pinto17a.html.

Stephen Prajna, Ali Jadbabaie, and George J. Pappas. A framework for worst-case and stochastic safety verification using barrier certificates. *IEEE Transactions on Automatic Control*, 52(8): 1415–1428, 2007. doi: 10.1109/TAC.2007.902736. URL https://doi.org/10.1109/TAC.2007.902736.

Marc Rigter, Bruno Lacerda, and Nick Hawes. RAMBO-RL: Robust adversarial model-based offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

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*, 2020. URL https://arxiv.org/abs/2004.03315.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Junwon Seo, Kensuke Nakamura, and Andrea Bajcsy. Uncertainty-aware latent safety filters for avoiding out-of-distribution failures. In *Conference on Robot Learning*, 2025. URL https://cmu-intentlab.github.io/UNISafe/.

Oswin So, Cheng Ge, and Chuchu Fan. Solving minimum-cost reach avoid using reinforcement learning. *Advances in Neural Information Processing Systems*, 37:30951–30984, 2024a.

Oswin So, Zachary Serlin, Makai Mann, Jake Gonzales, Kwesi Rutledge, Nicholas Roy, and Chuchu Fan. How to train your neural control barrier function: Learning safety filters for complex input-constrained systems. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11532–11539. IEEE, 2024b.

Oswin So, Eric Yu, and Chuchu Fan. Leap-cbf: A safety filter for uncertain systems with least-effort adversarial potentials. *arXiv preprint arXiv:2609.28364*, 2026.

Krishnan Srinivasan, Benjamin Eysenbach, Sehoon Ha, Jie Tan, and Chelsea Finn. Learning to be safe: Deep RL with a safety critic. *arXiv preprint arXiv:2010.14603*, 2020. URL https://arxiv.org/abs/2010.14603.

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*, 2021. URL https://arxiv.org/abs/2010.15920.

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.

Sander Tonkens and Sylvia Herbert. Refining control barrier functions through Hamilton-Jacobi reachability. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 13355–13362. IEEE, 2022.

Matteo Turchetta, Felix Berkenkamp, and Andreas Krause. Safe exploration in finite markov decision processes with gaussian processes. In *Advances in Neural Information Processing Systems*, 2016. URL https://arxiv.org/abs/1606.04753.

Kim P. Wabersich and Melanie N. Zeilinger. A predictive safety filter for learning-based control of constrained nonlinear dynamical systems. *Automatica*, 129:109597, 2021. doi: 10.1016/j.automatica.2021.109597. URL https://arxiv.org/abs/1812.05506.

He Yin, Andrew Packard, Murat Arcak, and Peter Seiler. Reachability analysis using dissipation inequalities for uncertain nonlinear systems. *Systems & Control Letters*, 142:104736, 2020. URL https://arxiv.org/abs/1808.02585.

Jun Zeng, Bike Zhang, and Koushil Sreenath. Safety-critical model predictive control with discrete-time control barrier function. In *American Control Conference (ACC)*, 2021.

Changjian Zhang, Parv Kapoor, Rômulo Meira-Góes, David Garlan, Eunsuk Kang, Akila Ganlath, Shatadal Mishra, and Nejib Ammar. Tolerance of reinforcement learning controllers against deviations in cyber physical systems. In *International Symposium on Formal Methods*, pp. 267–285. Springer, 2024. URL https://doi.org/10.1007/978-3-031-71177-0_17.

Hongchao Zhang, Junlin Wu, Yevgeniy Vorobeychik, and Andrew Clark. Exact verification of relu neural control barrier functions. *Advances in neural information processing systems*, 36:5685–5705, 2023.

Hongchao Zhang, Manan Tayal, Jackson Cox, Pushpak Jagtap, Shishir Kolathaya, and Andrew Clark. Stochastic neural control barrier functions. *arXiv preprint arXiv:2506.21697*, 2025a.

Songyuan Zhang, Oswin So, Mitchell Black, and Chuchu Fan. Discrete gcbf proximal policy optimization for multi-agent safe optimal control. In *International Conference on Learning Representations*, volume 2025, pp. 70839–70869, 2025b.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024.

A LATENT SAFETY TRANSFER

Let the physical state evolve as $s_{t+1} = f(s_t, a_t, w_t)$, where $f : \mathcal{S} \times \mathcal{A} \times \mathcal{W} \to \mathcal{S}$ and $w_t \in \mathcal{W}$ is a physical disturbance, with $0 \in \mathcal{W}$. As in the main text, observations are given by $o_t = o(s_t)$. Let $\mathcal{F} \subseteq \mathcal{S}$ denote the physical failure set and define the undisturbed reference transition $\bar{f}(s, a) \coloneqq f(s, a, 0)$. The physical disturbance w need not coincide with the world model's imagination noise.

During physical execution, the world model incorporates the next observation through a posterior update

$$z_{t+1} = g_{\text{post}}(z_t, o_{t+1}, a_t, \eta_t), \qquad \eta_t \sim q_{\text{post}}. \tag{15}$$

Here, $g_{\text{post}} : \mathcal{Z} \times \mathcal{O} \times \mathcal{A} \times \mathbb{R}^{n_\eta} \to \mathcal{Z}$ transforms noise from the base distribution q_{post} into posterior latent successors. Following the deterministic evaluation convention in (2), we define

$$\bar{g}_{\text{post}}(z, o, a) \coloneqq g_{\text{post}}(z, o, a, 0), \tag{16}$$

where we assume that $\eta = 0$ gives the mode of the stochastic posterior update.

We state a simple, sufficient condition for safety transfer in the following assumption.

Assumption 1 (Latent safety transfer)**.** There exists an encoder $E : \mathcal{S} \to \mathcal{Z}$ satisfying, for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$,

$$E(\bar{f}(s, a)) = \bar{g}(E(s), a) = \bar{g}_{\text{post}}(E(s), o(\bar{f}(s, a)), a), \tag{17a}$$

$$E(\mathcal{F}) \subseteq \hat{\mathcal{F}}. \tag{17b}$$

Under Assump. 1, the latent-admissible control guarantees the original undisturbed system remains outside the unsafe set $\mathcal{F}$. See App. B.1 for the formal statement and proof.

B PROOFS

B.1 D-CBF FORWARD INVARIANCE AND PHYSICAL SAFETY TRANSFER

We first state the physical-space counterpart of Def. 1. For a function $B : \mathcal{S} \to \mathbb{R}$, define $\mathcal{C}_B \coloneqq \{s \in \mathcal{S} : B(s) \geq 0\}$ and, for $s \in \mathcal{C}_B$,

$$\mathcal{A}_B(s) \coloneqq \{a \in \mathcal{A} : B(\bar{f}(s, a)) - B(s) \geq -\alpha(B(s))\}. \tag{18}$$

Here $\alpha : [0, \infty) \to [0, \infty)$ is a class-$\mathcal{K}$ function with $\alpha(r) \leq r$. The function B is a D-CBF for $\bar{f}$ relative to $\mathcal{F}$ if $\mathcal{C}_B \cap \mathcal{F} = \varnothing$ and such an α exists for which $\mathcal{A}_B(s)$ is nonempty at every $s \in \mathcal{C}_B$.

Theorem 2 (Forward invariance under D-CBF-admissible control)**.** *Let B be a D-CBF for $\bar{f}$ relative to $\mathcal{F}$, with associated function α and superlevel set $\mathcal{C}_B$. Then $\mathcal{A}_B(s)$ in* (18) *is nonempty for every $s \in \mathcal{C}_B$. Moreover, any control law selecting $a_t \in \mathcal{A}_B(s_t)$ whenever $s_t \in \mathcal{C}_B$ renders $\mathcal{C}_B$ forward invariant under $s_{t+1} = \bar{f}(s_t, a_t)$:*

$$s_0 \in \mathcal{C}_B \quad \Longrightarrow \quad s_t \in \mathcal{C}_B \subseteq \mathcal{S} \setminus \mathcal{F}, \qquad \forall t \geq 0. \tag{19}$$

Proof. Nonemptiness of $\mathcal{A}_B(s)$ for $s \in \mathcal{C}_B$ is part of the D-CBF condition. Now suppose $s_t \in \mathcal{C}_B$ and select any $a_t \in \mathcal{A}_B(s_t)$. Since $B(s_t) \geq 0$ and $\alpha(r) \leq r$ for every $r \geq 0$,

$$B(s_{t+1}) = B(\bar{f}(s_t, a_t)) \geq B(s_t) - \alpha(B(s_t)) \geq 0. \tag{20}$$

Thus $s_{t+1} \in \mathcal{C}_B$. Starting from $s_0 \in \mathcal{C}_B$, induction gives $s_t \in \mathcal{C}_B$ for all integers $t \geq 0$. Finally, $\mathcal{C}_B \cap \mathcal{F} = \varnothing$ implies $s_t \notin \mathcal{F}$. □

For the latent dynamics, define the corresponding admissible-action set for $z \in \hat{\mathcal{C}}$ as

$$\mathcal{A}_{\hat{B}}(z) \coloneqq \{a \in \mathcal{A} : \hat{B}(\bar{g}(z, a)) - \hat{B}(z) \geq -\alpha(\hat{B}(z))\}. \tag{21}$$

This set is nonempty by Def. 1. Applying Thm. 2 to $(\mathcal{Z}, \bar{g}, \hat{\mathcal{F}}, \hat{B})$ proves forward invariance of $\hat{\mathcal{C}}$ under the deterministic imagination dynamics and latent-admissible control. The following result states sufficient conditions for transferring this guarantee to physical execution with posterior updates.

Theorem 3 (Original-space safety)**.** *Let $\hat{B}$ be a D-CBF for $\bar{g}$ relative to $\hat{\mathcal{F}}$, and let $\pi_{\text{safe}} : \mathcal{Z} \to \mathcal{A}$ satisfy $\pi_{\text{safe}}(z) \in \mathcal{A}_{\hat{B}}(z)$ for every $z \in \hat{\mathcal{C}}$. Suppose there exists $E : \mathcal{S} \to \mathcal{Z}$ satisfying* (17a) *and*

(17b). *Then $B_{\text{phys}} \coloneqq \hat{B} \circ E$ is a D-CBF for $\bar{f}$ relative to $\mathcal{F}$. Moreover, the coupled dynamics*

$$a_t = \pi_{\text{safe}}(z_t), \qquad s_{t+1} = \bar{f}(s_t, a_t), \qquad z_{t+1} = \bar{g}_{\text{post}}(z_t, o(s_{t+1}), a_t) \tag{22}$$

with $z_0 = E(s_0) \in \hat{\mathcal{C}}$ satisfy

$$z_t = E(s_t) \in \hat{\mathcal{C}}, \qquad s_t \notin \mathcal{F}, \qquad \forall t \geq 0. \tag{23}$$

In particular, $E^{-1}(\hat{\mathcal{C}})$ is forward invariant under $s_{t+1} = \bar{f}(s_t, \pi_{\text{safe}}(E(s_t)))$ and is disjoint from $\mathcal{F}$.

Proof. The zero-superlevel set of B_{phys} is $E^{-1}(\hat{\mathcal{C}})$. If $s \in \mathcal{F}$, then (17b) gives $E(s) \in \hat{\mathcal{F}}$, hence $\hat{B}(E(s)) < 0$. Thus $E^{-1}(\hat{\mathcal{C}}) \cap \mathcal{F} = \varnothing$. For any $s \in E^{-1}(\hat{\mathcal{C}})$, consistency gives

$$B_{\text{phys}}(\bar{f}(s, a)) - B_{\text{phys}}(s) = \hat{B}(\bar{g}(E(s), a)) - \hat{B}(E(s)). \tag{24}$$

The admissible-action sets for B_{phys} at s and $\hat{B}$ at $E(s)$ therefore agree and are nonempty, with the same α. This proves the first claim.

We next show that the deployed latent state remains equal to the encoding of the physical state. If $z_t = E(s_t)$, then (17a) gives

$$z_{t+1} = \bar{g}_{\text{post}}(E(s_t), o(\bar{f}(s_t, a_t)), a_t) \tag{25}$$

$$= \bar{g}(E(s_t), a_t) = E(\bar{f}(s_t, a_t)) = E(s_{t+1}). \tag{26}$$

Since $z_0 = E(s_0)$, induction establishes $z_t = E(s_t)$ for all integers $t \geq 0$. The deployed latent trajectory thus follows the imagination dynamics with admissible actions, so latent forward invariance gives $z_t \in \hat{\mathcal{C}}$ for all $t \geq 0$. Consequently, $s_t \in E^{-1}(\hat{\mathcal{C}})$, which is disjoint from $\mathcal{F}$. Finally, $z_t = E(s_t)$ identifies the applied action with $\pi_{\text{safe}}(E(s_t))$, proving the claimed physical forward invariance. □

Remark 1 (Scope of the consistency conditions). The existence of $E : \mathcal{S} \to \mathcal{Z}$ is an idealized sufficient condition. A recurrent posterior tracker can assign different latent states to the same physical state after different observation-action histories, so such an encoder need not exist in general. Moreover, (17a) requires the deterministic imagined and posterior-updated successors to agree on encoded physical states; it does not cover arbitrary posterior corrections or stochastic posterior sampling. Equation (17b) permits conservative latent failure labels: a safe physical state may also be encoded into $\hat{\mathcal{F}}$, but a physical failure may not be encoded outside $\hat{\mathcal{F}}$.

B.2 Invariance to a constant margin shift

Proof of Lem. 4. For either Bellman backup T_h, translation commutes with expectation, maximum, and minimum, giving $T_{h+c}(V + c) = T_h(V) + c$. Thus value iteration initialized at the margin shifts by c throughout, while the maximizing actions remain unchanged because every action value shifts equally. □

B.3 Proofs for the minimum disturbance cost

For the one-step domination lemma, fix a policy π and abbreviate $J(z) = J^{\pi}_{\text{adv}}(z)$. Recall the conventions of Sec. 4.1: $\xi \geq 0$ with $\xi(0) = 0$, the infimum in Eq. (8) ranges over disturbance sequences whose closed-loop trajectory reaches $\hat{\mathcal{F}}$, and $\inf \emptyset = +\infty$, so $J : \mathcal{Z} \to [0, \infty]$.

Lem. 5 below is the dynamic-programming inequality underlying Lem. 1, obtained by prepending a single step to a feasible attack.

Lemma 5 (One-step domination)**.** *For every $z \notin \hat{\mathcal{F}}$ and every $\epsilon_0 \in \mathbb{R}^{n_\epsilon}$,*

$$J(z) \;\leq\; \xi(\epsilon_0) + J\big(g(z, \pi(z), \epsilon_0)\big). \tag{27}$$

Proof. Write $y \coloneqq g(z, \pi(z), \epsilon_0)$. If $J(y) = \infty$ the inequality is trivial, so assume $J(y) < \infty$ and let $\epsilon'_{0:\infty}$ be any disturbance sequence that is feasible from y, i.e. whose closed-loop trajectory from y reaches $\hat{\mathcal{F}}$, say at time $\tau' < \infty$, with cost $C = \sum_{t=0}^{\tau'-1} \xi(\epsilon'_t)$. Consider the prepended sequence $(\epsilon_0, \epsilon'_0, \epsilon'_1, \dots)$ started from z. Since $z \notin \hat{\mathcal{F}}$, its first step is taken, landing at $z_1 = y$; thereafter the

closed-loop trajectory coincides with the one from y under $\epsilon'_{0:\infty}$, so it reaches $\hat{\mathcal{F}}$ at time $1+\tau' < \infty$ with total cost $\xi(\epsilon_0) + C$. (If $y \in \hat{\mathcal{F}}$ the prepended sequence already fails at time 1 with cost $\xi(\epsilon_0)$, and $C \geq 0 = J(y)$ vacuously; the same bound holds.) Hence the prepended sequence is feasible from z and $J(z) \leq \xi(\epsilon_0)+C$. Taking the infimum over feasible $\epsilon'_{0:\infty}$ gives $J(z) \leq \xi(\epsilon_0)+J(y)$. □

For $z \in \hat{\mathcal{F}}$, the same one-step inequality holds because $J(z) = 0$ and both terms on the right are nonnegative. Thus, under the main text's finiteness assumption, rearranging the inequality with $\pi = \pi_{\mathrm{fb}}$ gives (9) at every state.

B.4 D-CBF PROPERTY OF THE SHIFTED MINIMUM DISTURBANCE COST

Proof of Lem. 1. The zero-superlevel set of $V_D = J^{\pi_{\mathrm{fb}}}_{\mathrm{adv}} - D$ is $\mathcal{S}_D$. Since $J^{\pi_{\mathrm{fb}}}_{\mathrm{adv}} = 0$ on $\hat{\mathcal{F}}$ and $D > 0$, this set is disjoint from $\hat{\mathcal{F}}$. Applying (9) with $\epsilon = 0$ and using $\xi(0) = 0$ gives, for every $z \in \mathcal{S}_D$,

$$V_D(\bar{g}(z, \pi_{\mathrm{fb}}(z))) - V_D(z) \geq 0 \geq -\alpha(V_D(z)). \tag{28}$$

The fallback therefore supplies an admissible action at every state in $\mathcal{S}_D$, establishing all conditions of Def. 1. □

B.5 DISTURBANCE-BUDGET CERTIFICATE UNDER THE FALLBACK

Proof of Lem. 2. Let $z_0 \in \mathcal{S}_D$ and suppose a disturbance sequence causes its first failure at time $\tau \leq t+1$ under π_{fb}. Since $D > 0$ and $J^{\pi_{\mathrm{fb}}}_{\mathrm{adv}} = 0$ on $\hat{\mathcal{F}}$, we have $\tau \geq 1$. This sequence is feasible in (8), so nonnegativity of the disturbance cost gives

$$D \leq J^{\pi_{\mathrm{fb}}}_{\mathrm{adv}}(z_0) \leq \sum_{k=0}^{\tau-1} \xi(\epsilon_k) \leq c(\epsilon_{0:t}). \tag{29}$$

Thus, $c(\epsilon_{0:t}) < D$ excludes failure through time $t+1$. □

B.6 DISTURBANCE-BUDGET CERTIFICATE UNDER THE SAFETY FILTER

Proof of Lem. 3. Define $K(z) := \min\{J^{\pi_{\mathrm{fb}}}_{\mathrm{adv}}(z), D\}$. Since $\alpha(r) \leq r$ for $r \geq 0$, admissibility in (12) implies

$$Q(z, \pi(z)) \geq J^{\pi_{\mathrm{fb}}}_{\mathrm{adv}}(z) - \alpha(\max\{J^{\pi_{\mathrm{fb}}}_{\mathrm{adv}}(z) - D, 0\}) \geq K(z). \tag{30}$$

Thus, by (11), every step before the first failure satisfies

$$\xi(\epsilon_t) + J^{\pi_{\mathrm{fb}}}_{\mathrm{adv}}(z_{t+1}) \geq K(z_t). \tag{31}$$

As $K(z_t) \leq D$ and $\xi(\epsilon_t) \geq 0$, capping the next value at D preserves this inequality: $\xi(\epsilon_t) + K(z_{t+1}) \geq K(z_t)$. If the trajectory reaches failure at time τ, telescoping gives

$$c(\epsilon_{0:\tau-1}) + K(z_\tau) \geq K(z_0) = D. \tag{32}$$

Since $J^{\pi_{\mathrm{fb}}}_{\mathrm{adv}}$ is zero on $\hat{\mathcal{F}}$, $K(z_\tau) = 0$, so failure requires cumulative disturbance cost at least D. Nonnegativity of ξ then proves the stated implication for every $t \geq 0$. Finally, (9) gives $Q(z, \pi_{\mathrm{fb}}(z)) \geq J^{\pi_{\mathrm{fb}}}_{\mathrm{adv}}(z)$, so the fallback action satisfies (12). □

B.7 FINITE-HORIZON FAILURE PROBABILITY

Proof of Thm. 1. Since $z_0 \in \mathcal{S}_D$ and $D > 0$, the initial state is outside $\hat{\mathcal{F}}$. On the event $\{\tau \leq N\}$, Lem. 3 implies $c(\epsilon_{0:\tau-1}) \geq D$. Nonnegativity of ξ then gives

$$\{\tau \leq N\} \subseteq \{c(\epsilon_{0:N-1}) \geq D\}. \tag{33}$$

Taking probabilities proves the claim. For $N = 0$, both events are empty, with the empty cost sum equal to zero. □

B.8 Gaussian Disturbance Budget Bound

For a noise density q with a mode at zero, one possible disturbance cost is $\xi(\epsilon) = \log q(0) - \log q(\epsilon)$ wherever $0 < q(\epsilon) \le q(0) < \infty$. This cost is nonnegative, vanishes at zero, and assigns greater effort to lower-density realizations. For standard Gaussian noise, it reduces to $\xi(\epsilon) = \frac{1}{2}\|\epsilon\|^2$.

Corollary 2. *Under the admissible policy of Thm. 1, for independent standard normal disturbances $\epsilon_t \sim \mathcal{N}(0, I_{n_\epsilon})$ and $\xi(\epsilon) = \frac{1}{2}\|\epsilon\|^2$, $2c(\epsilon_{0:N-1})$ follows a χ^2 distribution with Nn_ϵ degrees of freedom. Thus,*

$$\mathbb{P}\Big(\tau \le N\Big) \le \mathbb{P}\Big(\chi^2_{Nn_\epsilon} \ge 2D\Big), \qquad 0 < D \le J^{\pi_{\mathrm{fb}}}_{\mathrm{adv}}(z_0). \tag{34}$$

Consequently, for any target violation probability $\delta \in (0,1)$, it is sufficient to choose D such that

$$\frac{1}{2} F^{-1}_{\chi^2_{Nn_\epsilon}}(1-\delta) \le D \le J^{\pi_{\mathrm{fb}}}_{\mathrm{adv}}(z_0), \tag{35}$$

where $F^{-1}_{\chi^2_{Nn_\epsilon}}$ is the inverse CDF of the χ^2 distribution with Nn_ϵ degrees of freedom.

Proof of Cor. 1 and the budget quantile condition. For a positive integer horizon N, independence across time and the standard normal law of each disturbance give

$$2c(\epsilon_{0:N-1}) = \sum_{t=0}^{N-1} \sum_{j=1}^{n_\epsilon} \epsilon_{t,j}^2 \sim \chi^2_{Nn_\epsilon}. \tag{36}$$

By Thm. 1 and continuity of the χ^2 distribution,

$$\mathbb{P}(\tau \le N) \le \mathbb{P}(2c(\epsilon_{0:N-1}) \ge 2D) = 1 - F_{\chi^2}(2D; Nn_\epsilon). \tag{37}$$

This proves Cor. 1. The right-hand side is at most δ whenever $2D \ge F^{-1}_{\chi^2_{Nn_\epsilon}}(1-\delta)$. Combining this condition with $D \le J^{\pi_{\mathrm{fb}}}_{\mathrm{adv}}(z_0)$ gives the stated sufficient budget interval. □

B.9 Extended-Valued Minimum Disturbance Costs

The definition in (8) permits $J^{\pi_{\mathrm{fb}}}_{\mathrm{adv}}(z) = +\infty$ when no finite-cost disturbance sequence reaches failure. To obtain a real-valued barrier without assuming finiteness everywhere, fix a finite cap $M > D > 0$ and define

$$J_M(z) \coloneqq \min\{J^{\pi_{\mathrm{fb}}}_{\mathrm{adv}}(z), M\}, \qquad V_{D,M}(z) \coloneqq J_M(z) - D. \tag{38}$$

The cap preserves the one-step inequality needed for safety filtering. Indeed, writing $y = g(z, \pi_{\mathrm{fb}}(z), \epsilon)$, the extended-valued inequality $J^{\pi_{\mathrm{fb}}}_{\mathrm{adv}}(z) \le \xi(\epsilon) + J^{\pi_{\mathrm{fb}}}_{\mathrm{adv}}(y)$ implies

$$J_M(z) \le \min\{\xi(\epsilon) + J^{\pi_{\mathrm{fb}}}_{\mathrm{adv}}(y), M\} \le \xi(\epsilon) + J_M(y), \tag{39}$$

where the second inequality uses $\xi(\epsilon) \ge 0$. Thus, J_M cannot decrease under the zero-noise fallback transition, since $\xi(0) = 0$. Also, $J_M = 0$ on $\hat{\mathcal{F}}$, and $\{z : J_M(z) \ge D\} = \mathcal{S}_D$. Consequently, $V_{D,M}$ satisfies the same D-CBF condition as the finite-valued construction.

For filtering, replace Q by the finite action value

$$Q_M(z,a) \coloneqq \inf_{\epsilon \in \mathbb{R}^{n_\epsilon}} \{\xi(\epsilon) + J_M(g(z,a,\epsilon))\} \tag{40}$$

and impose

$$Q_M(z,a) - J_M(z) \ge -\alpha(\max\{J_M(z) - D, 0\}). \tag{41}$$

The fallback remains admissible by the capped one-step inequality. The proof of Lem. 3 applies with J_M in place of $J^{\pi_{\mathrm{fb}}}_{\mathrm{adv}}$: the reserve $K(z) = \min\{J_M(z), D\}$ satisfies $\xi(\epsilon_t) + K(z_{t+1}) \ge K(z_t)$ before failure. Starting in $\mathcal{S}_D$, telescoping to a failure state therefore requires cumulative cost at least D. The probability bound in Thm. 1 follows unchanged for this capped filter. This construction preserves the budget certificate and its initial set, although capping can change which actions the filter admits.

B.10 Expectation-based Discrete-time Control Barrier Functions

HJ reachability with a stochastic backup (e.g., the one used in (Nakamura et al., 2026)) solves for the following problem:

$$V_{\mathrm{HJ}}(s) = \min\Big\{h(s), \max_{a\in\mathcal{A}} \mathbb{E}_{\epsilon\sim q}[V_{\mathrm{HJ}}(g(s,a,\epsilon))]\Big\}, \tag{42}$$

where g is the stochastic imagination dynamics (1) and q is the distribution of the stochasticity ϵ.

Theorem 4. *V_{HJ} satisfies the following expectation-based D-CBF condition:*

$$\max_{a\in\mathcal{A}} \mathbb{E}_{\epsilon\sim q}[V_{\mathrm{HJ}}(g(s,a,\epsilon))] - V_{\mathrm{HJ}}(s) \geq 0 \geq -\alpha(V_{\mathrm{HJ}}(s)), \quad \forall s \in \mathcal{C}_{V_{\mathrm{HJ}}}, \tag{43}$$

for class-$\mathcal{K}$ function α with $\alpha(r) \leq r$ for $r \geq 0$.

Proof. By the definition of $V_{\mathrm{HJ}}(s)$, we have:

$$\begin{aligned} V_{\mathrm{HJ}}(s) &= \min\Big\{h(s), \max_{a\in\mathcal{A}} \mathbb{E}_{\epsilon\sim q}[V_{\mathrm{HJ}}(g(s,a,\epsilon))]\Big\} \\ &\leq \max_{a\in\mathcal{A}} \mathbb{E}_{\epsilon\sim q}[V_{\mathrm{HJ}}(g(s,a,\epsilon))]. \end{aligned}$$

For $s \in \mathcal{C}_{V_{\mathrm{HJ}}}$, we have $V_{\mathrm{HJ}}(s) \geq 0$ and thus $-\alpha(V_{\mathrm{HJ}}(s)) \leq 0$. Therefore,

$$\max_{a\in\mathcal{A}} \mathbb{E}_{\epsilon\sim q}[V_{\mathrm{HJ}}(g(s,a,\epsilon))] - V_{\mathrm{HJ}}(s) \geq 0 \geq -\alpha(V_{\mathrm{HJ}}(s)). \qquad \square$$

Let $Q_{\mathrm{HJ}} : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ be the Q-function defined as

$$Q_{\mathrm{HJ}}(s,a) = \mathbb{E}_{\epsilon\sim q}[V_{\mathrm{HJ}}(g(s,a,\epsilon))]. \tag{44}$$

We can then define the admissible control set $\mathcal{A}_{V_{\mathrm{HJ}}}$ for the stochastic safety filter as

$$\mathcal{A}_{V_{\mathrm{HJ}}}(s) = \{a \in \mathcal{A} : Q_{\mathrm{HJ}}(s,a) - V_{\mathrm{HJ}}(s) \geq -\alpha(V_{\mathrm{HJ}}(s))\}. \tag{45}$$

Given a nominal policy π_{nom}, V_{HJ} can be used to construct the following safety filter:

$$\pi_{\mathrm{safe}}(s) = \operatorname*{arg\,min}_{a\in\mathcal{A}_{V_{\mathrm{HJ}}}(s)} \|a - \pi_{\mathrm{nom}}(s)\|. \tag{46}$$

We then have the following finite-horizon safety guarantee for the stochastic imagination dynamics inspired by the bound from Cosner et al. (2023, Theorem 5).

Theorem 5 (Safety Guarantee for Stochastic D-CBF)**.** *Suppose V_{HJ} is bounded above by $M > 0$ and $\alpha(r) \leq \beta r$ for all $r \in [0, M]$, where $\beta \in (0,1)$. Assume the safety filter admits a measurable minimizing selection on $\mathcal{C}_{V_{\mathrm{HJ}}}$, the expectations defining $Q_{\mathrm{HJ}}(s, \pi_{\mathrm{safe}}(s))$ are finite there, and the transition noises are independent draws from q. Then, for any initial state $s_0 \in \mathcal{C}_{V_{\mathrm{HJ}}}$, let τ be the first time step that the system leaves the safe set $\mathcal{C}_{V_{\mathrm{HJ}}}$ under the stochastic imagination dynamics g and the safety filter π_{safe}, i.e.,*

$$\tau = \inf\{k \geq 0 : s_k \notin \mathcal{C}_{V_{\mathrm{HJ}}}\}. \tag{47}$$

Then, for every integer $N \geq 0$,

$$\mathbb{P}\Big(\tau \leq N\Big) \leq 1 - \frac{V_{\mathrm{HJ}}(s_0)}{M}(1-\beta)^N. \tag{48}$$

Proof. Let $V_k = V_{\mathrm{HJ}}(s_k)$ and $\mathcal{F}_k = \sigma(s_0, \epsilon_0, \ldots, \epsilon_{k-1})$. For this argument, stop the trajectory at τ; its evolution after the first exit does not affect the exit probability. On $\{\tau > k\}$, admissibility of $\pi_{\mathrm{safe}}(s_k)$ and independence of the next noise give

$$\begin{aligned} \mathbb{E}[V_{k+1} \mid \mathcal{F}_k] &= Q_{\mathrm{HJ}}(s_k, \pi_{\mathrm{safe}}(s_k)) \\ &\geq V_k - \alpha(V_k) \geq (1-\beta)V_k. \end{aligned}$$

Define $Y_k = V_k \mathbf{1}_{\{\tau>k\}}$, so that $0 \leq Y_k \leq M$. Since $\{\tau > k+1\} = \{\tau > k\} \cap \{V_{k+1} \geq 0\}$, discarding the $\{V_{k+1} < 0\}$ part gives

$$\begin{aligned} \mathbb{E}[Y_{k+1} \mid \mathcal{F}_k] &= \mathbf{1}_{\{\tau>k\}} \mathbb{E}[V_{k+1}\mathbf{1}_{\{V_{k+1}\geq 0\}} \mid \mathcal{F}_k] \\ &\geq \mathbf{1}_{\{\tau>k\}} \mathbb{E}[V_{k+1} \mid \mathcal{F}_k] \\ &\geq (1-\beta)Y_k. \end{aligned}$$

Taking expectations and iterating, with $Y_0 = V_{\mathrm{HJ}}(s_0)$, gives

$$(1-\beta)^N V_{\mathrm{HJ}}(s_0) \le \mathbb{E}[Y_N] \le M\mathbb{P}(\tau > N).$$

Dividing by M then taking complements yields the result. □

B.11 Toy Example

Dynamics and common nominal policy. Let $x_0 = 4$, $u \in \{0,1,2\}$, and $\epsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. The dynamics, failure set, and first failure time are

$$x^+ = g(x,u,\epsilon) = \begin{cases} x, & u=0 \quad \text{(hold)}, \\ \epsilon, & u=1 \quad \text{(conservative reset)}, \\ 4\epsilon, & u=2 \quad \text{(risky reset)}, \end{cases} \qquad \mathcal{F} = \{x : |x| > 5\}, \qquad \tau = \inf\{t \ge 0 : x_t \in \mathcal{F}\}. \tag{49}$$

Both filters minimize $(u - \pi_{\mathrm{nom}}(x))^2$ over their admissible actions, where $\pi_{\mathrm{nom}}(x) = 0$ for $|x| < 1$ and $\pi_{\mathrm{nom}}(x) = 2$ otherwise. Thus both hold forever after reaching the core $(-1,1)$; outside it they prefer risky reset, then conservative reset, then hold. The plotted quantity is the actual safety probability $\mathbb{P}(\tau > N)$.

Exact stochastic HJ value. Choose $h(x) = 25 - x^2$. The expected successor values of h are $h(x)$, 24, and 9 for actions 0, 1, and 2, respectively. Hence

$$\min\{h(x), \max_u \mathbb{E}[h(g(x,u,\epsilon))]\} = \min\{h(x), \max\{h(x), 24, 9\}\} = h(x). \tag{50}$$

Thus $V_{\mathrm{HJ}} = h$ is an exact Bellman fixed point, and is the greatest fixed point because every fixed point is bounded above by h. Bellman iteration initialized at h leaves it unchanged. The upper bound $M = 25$ is exact and the relevant expectations are finite.

Safest stochastic D-CBF relaxation. The admissibility condition is $Q_{\mathrm{HJ}}(x,u) \ge V_{\mathrm{HJ}}(x) - \alpha(V_{\mathrm{HJ}}(x))$. In the zero-relaxation limit it yields

$$\pi_0(x) = \begin{cases} 0, & |x| < 1, \\ 1, & 1 \le |x| < 4, \\ 2, & 4 \le |x| \le 5. \end{cases} \tag{51}$$

In particular $Q_{\mathrm{HJ}}(4,2) = 9 = V_{\mathrm{HJ}}(4)$, so every nonnegative relaxation accepts the risky nominal action initially. Its failure probability is $p_4 = 2\Phi(-5/4) \approx 0.21129955$, where Φ is the standard normal CDF. The zero function is not strictly class-$\mathcal{K}$: the plotted policy is the limit of $\alpha_\beta(r) = \beta r$ as $\beta \downarrow 0$.

This limit minimizes actual failure probability over all admissible relaxations, with h and the nominal objective fixed. Indeed, enlarging the admissible action set can only increase the selected action index outside the core. Under the limiting policy, failure-to-go is zero in the core, constant in each reset region, and nondecreasing in $|x|$. For any such function f, coupling the resets gives $\mathbb{E}[f(\epsilon)] \le \mathbb{E}[f(4\epsilon)]$. Induction on the remaining horizon therefore proves optimality of the smaller feasible action index selected by π_0. Positive linear relaxations approach this policy and its finite-horizon probabilities continuously. This optimizes the relaxation, not the choice of h or the nominal objective; arbitrary admissible-action selection could always choose hold and avoid failure.

Minimum-effort certificate and budget regimes. Use fallback $\pi_{\mathrm{fb}}(x) = 1$ and effort $\ell(e) = e^2/2$. Failure under this fallback requires at least one $|e| > 5$. A single disturbance tending to 5 from above shows that the exact minimum-effort infimum is

$$J(x) = \begin{cases} 25/2, & |x| \le 5, \\ 0, & |x| > 5. \end{cases} \qquad Q_J(x,u) = \inf_e\{e^2/2 + J(g(x,u,e))\}. \tag{52}$$

At every safe state, $Q_J(x,0) = Q_J(x,1) = 25/2$ and $Q_J(x,2) = 25/32$. For the WISH toy-example analysis, we specialize Eq. (12) to $\alpha(r) = r$, so that

$$J(x) - \alpha(\max\{J(x) - D, 0\}) = \min\{J(x), D\}.$$

The resulting guard $Q_J(x,u) \geq \min\{J(x), D\}$ is initially certified for $0 < D \leq 25/2$. It selects hold in the core, risky reset outside it when $0 < D \leq 25/32$, and conservative reset outside it when $25/32 < D \leq 25/2$. Equality at the threshold admits risky reset.

Exact probabilities. Let $\overline{\Phi} = 1 - \Phi$ and define, for $\sigma \in \{1, 4\}$,

$$A_\sigma = 1 - 2\overline{\Phi}(1/\sigma), \qquad B_\sigma = 2[\overline{\Phi}(1/\sigma) - \overline{\Phi}(4/\sigma)], \tag{53}$$

$$C_\sigma = 2[\overline{\Phi}(4/\sigma) - \overline{\Phi}(5/\sigma)], \qquad p_\sigma = 2\overline{\Phi}(5/\sigma). \tag{54}$$

These are the reset probabilities of entering the hold, conservative, risky, and failure regions of (51). With $Q = \left(\begin{smallmatrix} B_1 & C_1 \\ B_4 & C_4 \end{smallmatrix}\right)$, $p = (p_1, p_4)^\top$, and $e_R = (0,1)^\top$, the limiting D-CBF risk is

$$F^*_{\mathrm{HJ}}(N) = e_R^\top \sum_{j=0}^{N-1} Q^j p = e_R^\top (I - Q^N)(I - Q)^{-1} p. \tag{55}$$

For the budget filter with $\alpha(r) = r$, let $r_\sigma = 1 - A_\sigma - p_\sigma$. Then

$$F_D(N) = \frac{p_\sigma}{A_\sigma + p_\sigma}(1 - r_\sigma^N), \qquad \sigma = \begin{cases} 4, & 0 < D \leq 25/32, \\ 1, & 25/32 < D \leq 25/2. \end{cases} \tag{56}$$

In each case the safety probability is $1 - F(N)$. At $N = 10$, the actual risks are approximately 0.23636811 for the safest D-CBF limit, 0.51428794 for the permissive budget, and 8.3976×10^{-7} for the protective budget. The protective-budget advantage holds at every positive horizon: both policies hold in the core, while elsewhere it always uses reset scale 1 and the D-CBF sometimes uses scale 4. Applying the same monotonicity argument to the protective-budget failure-to-go proves the ordering, with strictness from the initial risky D-CBF action. Their limiting risks are approximately 8.3977×10^{-7} and 0.23636811, respectively.

C Experimental Details

C.1 Tasks and Evaluation Metrics

We evaluate on four continuous-control tasks: Cartpole, Cheetah, Walker, and Waiter. The first three are MuJoCo environments from the DeepMind Control Suite; Waiter is our custom MuJoCo (Todorov et al., 2012) manipulation task. Each task uses a separately trained, frozen DreamerV3 world model. The recurrent state has a 32-dimensional stochastic component and a 512-dimensional deterministic component, whose concatenation is the 544-dimensional feature used by the learned values and policies. Observations are 64×64 RGB images, and one learned transition is one RSSM imagination step.

The privileged simulator states described below are used only to generate failure labels and compute evaluation metrics. The deployed filters never observe them: they update the RSSM state from images and act through learned latent failure heads. In the reported WISH runs, the adversary perturbs only the 32-dimensional stochastic RSSM innovation,

$$z'_{\mathrm{stoch}} = \mu(z,u) + \sigma(z,u) \odot \epsilon, \qquad \epsilon \in [-1,1]^{32},$$

and pays $\frac{1}{2}\|\epsilon\|_2^2$ per step; it does not perturb the physical action channel. During fallback training, each imagined step uses a $50/50$ mixture of a Gaussian-prior innovation and an adversarial innovation. At evaluation time, all task metrics are truncated at the first failure or the evaluation horizon.

World-model data. For Cartpole, Cheetah, and Walker, data collection uses visible OU action exploration at scales 0, 0.2, and 0.5. Half of the episodes additionally receive unobserved white actuation noise at the task's reported evaluation scale, and the world model receives the commanded action. These models are trained for $60{,}000$, $90{,}000$, and $90{,}000$ updates, respectively. The Waiter model is trained for $120{,}000$ updates on visible OU exploration at the same three scales, without hidden actuation noise, and receives the executed action. Its collection environment uses a $0.1\,\mathrm{m}$ soft workspace edge, disables cast shadows, and hides the arm from the world-model camera. The evaluation noise standard deviations are 1.0, 0.5, 0.2, and 1.0 for Cartpole, Cheetah, Walker, and Waiter, respectively; the corresponding horizons are 300, 300, 300, and 500 steps.

C.1.1 CARTPOLE SWING-UP

The agent swings a pole upright while keeping its cart on a finite rail. The privileged state is $(x, \theta, \dot{x}, \dot{\theta}) \in \mathbb{R}^4$, the one-dimensional action is the normalized horizontal force on the cart, and the control time step is $\Delta t = 0.02$ seconds. Failure occurs when the cart leaves the central portion of the rail,

$$\mathcal{A}_{\text{cartpole}} = \{\, |x| > 1.26 \,\}.$$

Performance is upright time, or the number of pre-failure steps for which $\cos(\theta) > 0.9$, multiplied by Δt.

C.1.2 CHEETAH RUN

In Cheetah, the agent controls six joint torques. The privileged state contains nine generalized positions and nine generalized velocities, and the control time step is $\Delta t = 0.02$ seconds. Let θ_{torso} be the wrapped root pitch. Failure occurs once the torso tips more than 50° from its nominal orientation,

$$\mathcal{A}_{\text{cheetah}} = \{\, |\,\text{wrap}(\theta_{\text{torso}})| > 50^\circ \,\}.$$

Performance is forward distance $x_T - x_0$ in meters.

C.1.3 WALKER RUN

In Walker, the agent controls six joint torques. The privileged state again contains nine generalized positions and nine generalized velocities, and the control time step is $\Delta t = 0.05$ seconds. Evaluation starts from a standing configuration with zero velocity and a torso-pitch perturbation sampled uniformly from $[-0.1, 0.1]$ radians. Failure occurs when the torso reaches the ground,

$$\mathcal{A}_{\text{walker}} = \{\, h_{\text{torso}} < 0.37 \text{ m} \,\}.$$

Performance is forward distance $x_T - x_0$ in meters.

C.1.4 WAITER

In this custom task, a Franka Panda arm carries a loose bottle on a tray through a sequence of randomly sampled planar targets. The four-dimensional normalized command specifies the tray's horizontal x–y velocity and its roll and pitch rates. Damped least-squares inverse kinematics maps this command to the seven joint-position actuators, and the control time step is $\Delta t = 0.02$ seconds.

Let (x_o, y_o, z_o) be the bottle-base position in the tray frame, $r_o = \sqrt{x_o^2 + y_o^2}$, and θ_o the angle between the bottle axis and tray normal. Failure occurs when the bottle slides beyond the tray radius, drops below the tray, or tips too far:

$$\mathcal{A}_{\text{waiter}} = \{\, r_o > 0.16 \text{ m} \,\} \cup \{\, z_o < -0.02 \text{ m} \,\} \cup \{\, \theta_o > 50^\circ \,\}.$$

A delivery is completed when the tray is within 0.04 meters of its target and moves slower than 0.3 meters per second, after which a new target is sampled. Performance is completed deliveries plus partial credit for the current target leg,

$$N_{\text{delivered}} + \text{clip}_{[0,1]}\left(1 - \frac{\|p_{\text{tray}} - p_{\text{target}}\|_2}{\ell_{\text{leg}}}\right),$$

where ℓ_{leg} is the initial distance to that target.

C.2 BASELINES

We compare WISH against one unfiltered controller and three safety filters. All four run the same nominal task policy, the same world model, and (for the filtered arms) the same failure detector, so every gap is attributable to the safety mechanism rather than the underlying policy or model.

Nominal. The nominal task policy π_{nom} with no safety filter.

HJ (no GP) (Nakamura et al., 2025). A learned Hamilton-Jacobi (HJ) reachability filter in latent space using a binary classifier margin head. The failure set is specified with binary fail labels on

observations, and the learned margin function is a binary classifier fit to those labels. As Nakamura et al. (2026) show, a binary classification margin yields an HJ value function that gives little gradient information across actions for safety filtering.

HJ (GP) (Nakamura et al., 2026). The same HJ reachability filter as Nakamura et al. (2025), but with a WGAN-GP regularized *smooth* margin trained from the same binary labels. This yields a non-saturated margin whose HJ value varies more smoothly across the boundary and thus gives a usable filtering signal.

MPS (Bastani, 2021). A model-predictive shielding filter. Rather than learning a value function, it forward-simulates H steps in the world model with a single nominal step first, then a provided fallback π_{fb} for the remaining horizon. If the fallback fails in imagination within H steps, the filter overrides to the fallback control.

C.3 Implementation Details

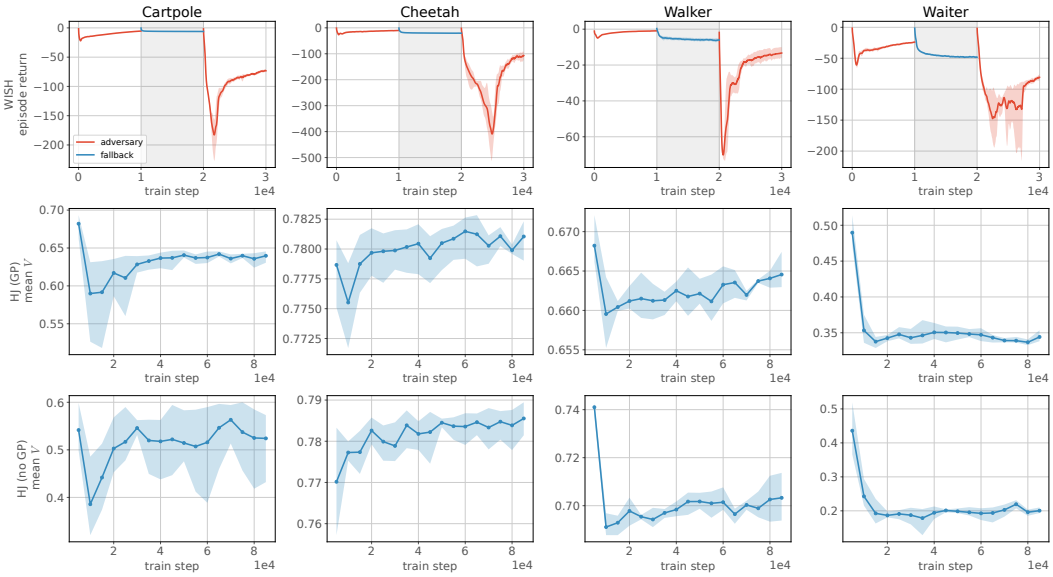

Figure 6: **Training curves for the world-model tasks.** Columns correspond to Cartpole, Cheetah, Walker, and Waiter. The top row reports imagined-rollout episode returns during alternating WISH adversary and fallback phases; gray intervals denote fallback-training phases. The middle and bottom rows report the mean value $\hat{V}$ during HJ (GP) and HJ (no GP) training, respectively. Solid curves are means over three training seeds and shaded bands span the minimum and maximum across seeds. The horizontal axes use method-specific training units: cumulative PPO iterations for WISH and actor-critic gradient updates after warm-up for HJ.

Tab. 1 lists the training hyperparameters of WISH and the HJ baselines. All four world-model tasks share them, except where the table notes a per-task value, and every training seed uses the same configuration. HJ (GP) and HJ (no GP) differ only in the failure-margin head of the world model they train on (with or without a gradient penalty).

Table 1: **Training hyperparameters** for Cartpole, Cheetah, Walker, and Waiter.

Hyperparameter	Value
WISH (PPO, in world-model imagination)	
Parallel imagined rollouts × rollout length	1024×15
Start-latent pool (refreshed every 25 iterations)	16384
Training schedule (PPO iterations)	adversary 10^4, fallback 10^4, adversary 10^4
Final critic polish (iterations)	4000
Discount γ (adversary / fallback)	0.999 / 0.99
GAE λ	0.95
PPO clip ϵ	0.2
Epochs per iteration / minibatch size	1 / 4096
Entropy coefficient (linearly annealed)	$0.01 \to 0.005$
Actor / critic learning rate (Adam)	10^{-4} / 5×10^{-4}
Max gradient norm	0.5
Actor / critic network	MLP, 2×256
Adversary channel	32-D RSSM innovation only (no action push)
Latent innovation bound	$\|\epsilon\|_\infty \le 1.0$ (cube)
Latent effort	$\frac{1}{2}\|\epsilon\|_2^2$ per step (weight 1.0)
Latent adversary initial std	0.1
Fallback training dynamics	per step, 50/50 adversarial / Gaussian-prior innovation
Safety classifier: network, learning rate, batch	MLP 128, 10^{-3}, 4096
Safety classifier: discount, fit steps	0.99, 4000
Reach classifier: network, learning rate, batch	MLP 128, 10^{-3}, 4096
Reach classifier: label rollout, label discount	300 steps, 0.99966
Reach classifier: fit steps, critic split threshold	3000, 0.5
Time step Δt	0.02 (0.05 on Walker)
HJ (GP) / HJ (no GP) (DDPG-style actor–critic, in world-model imagination)	
Epochs	1 warm-up +16
Imagined episodes per epoch × length	5000×8 steps
Gradient updates	1 per episode (85000 total)
Episode start	3-frame window from the world-model dataset
Episodes replaying the nominal policy	50%
Replay buffer / batch size	40000 / 512
Discount γ	0.9999
Actor learning rate / weight decay (AdamW)	10^{-4} / 0.01
Critic learning rate / weight decay (AdamW)	10^{-3} / 10^{-3}
Target-network update rate τ	0.005
Exploration noise std	0.1
Stage-cost transform	tanh
Actor / critic network	MLP, 2×256

C.3.1 Classifier fitting and policy evaluation

After the final adversary refit, we fit two classifiers with the policies frozen. The reachability classifier $c_{\text{reach}} : \mathcal{Z} \to [0,1]$ predicts whether deterministic π_{adv} can drive π_{fb} into $\hat{\mathcal{F}}$. We train it from 300-step rollouts of $(\pi_{\text{adv}}, \pi_{\text{fb}})$, using discounted bootstrap targets when a rollout does not reach failure. The adversary has feasible and infeasible value heads, trained on states with $c_{\text{reach}}(z) \ge 0.5$ and $c_{\text{reach}}(z) < 0.5$, respectively. Writing $\widehat{J}_{\text{feas}} = -V_{\text{adv}}^{\text{feas}}$ and $\widehat{J}_{\text{infeas}} = -V_{\text{adv}}^{\text{infeas}}$, the deployed estimate is the soft composition

$$\widehat{J}_{\text{dep}}(z) = c_{\text{reach}}(z)\widehat{J}_{\text{feas}}(z) + \big(1 - c_{\text{reach}}(z)\big)\widehat{J}_{\text{infeas}}(z). \tag{57}$$

The fallback-safety score $p_{\text{safe}} : \mathcal{Z} \to [0,1]$ is the complement of the estimated probability that undisturbed π_{fb} reaches failure, so larger values indicate safer states; it is fit from 15-step rollouts followed by a discounted bootstrap. We then perform 4000 fixed-policy- evaluation iterations with both policies frozen, deterministic π_{adv}, and only $V_{\text{adv}}^{\text{feas}}$ updated; $V_{\text{adv}}^{\text{infeas}}$ remains frozen to preserve its exploration-trained scale. The polish uses $\hat{\mathcal{F}}_{\text{polish}} = \hat{\mathcal{F}} \cup \{z \in \mathcal{Z} : p_{\text{safe}}(z) \le 0.5\}$, making states predicted unsafe for π_{fb} zero-return and absorbing. The score is retained for training and diagnostics but does not mask $\widehat{J}_{\text{dep}}$ at deployment.

C.3.2 DEPLOYMENT-TIME FILTERING

In the implementation, we specialize to the linear function $\alpha(r) = \alpha r$, abusing notation by using $\alpha \in (0, 1]$ also for its scalar coefficient. At each step, the WISH filter evaluates the nominal action and 25 evenly spaced actions along the line segment from the nominal action to the learned fallback action. For each candidate action, we approximate the infimum in Eq. (11) with 64 one-step latent disturbances: 32 samples from the standard-normal RSSM prior and 32 samples from the learned adversary. The candidate score is the minimum of $\frac{1}{2}\|\epsilon\|_2^2 + \widehat{J}_{\mathrm{dep}}(z')$ over these samples. Writing this score as $\widehat{Q}(z, a)$, the implementation accepts a candidate only if $\widehat{Q}(z, a) - D \geq (1-\alpha)(\widehat{J}_{\mathrm{dep}}(z) - D)$ and $\widehat{Q}(z, a) \geq D$. We execute the nominal action if it is accepted; otherwise, we execute the closest accepted candidate, or the pure fallback action if no candidate is accepted. The second inequality makes the sampled implementation stricter than Eq. (12) when $\widehat{J}_{\mathrm{dep}}(z) < D$.

C.3.3 EVALUATION PROTOCOL

We evaluate every method closed-loop in the original simulator, using the frozen world model only inside the deployed safety filter. At each agent step, we perturb the normalized action with independent white Gaussian noise, $a_t^{\mathrm{exec}} = \mathrm{clip}_{[-1,1]}(a_t + \sigma\xi_t)$ with $\xi_t \sim \mathcal{N}(0, I)$. In normalized action units, we use $\sigma = 1.0$, 0.5, 0.2, and 1.0 for Cartpole, Cheetah, Walker, and Waiter, respectively. Episodes last 300 steps for the first three tasks and 500 steps for Waiter. For each filter setting, we run 40 episodes for every combination of three training seeds and three evaluation seeds, for $9 \times 40 = 360$ episodes per operating point; methods evaluated under the same evaluation seed receive matched initial conditions and disturbance sequences. Safety rate is the fraction of episodes that never enter the failure set before the evaluation horizon, and task performance is computed on the trajectory truncated at the first failure or the evaluation horizon. For WISH and both HJ baselines, we sweep 20 robustness thresholds and $\alpha \in \{0.1, 0.5, 0.8\}$; for MPS, we sweep 20 margin thresholds with rollout horizon $H = 5$. We average each setting across the nine seed pairs before computing its safety-performance Pareto frontier.

D EXTENDED RELATED WORK

WISH connects safety filtering, learning in latent dynamics, and adversarial failure search. The comparisons below distinguish the information used to specify safety, the quantity a learned value represents, and the assumptions that support deployment guarantees. These distinctions matter because failure membership, expected constraint cost, and resistance to disturbances answer different safety questions.

D.1 REACHABILITY, VIABILITY, AND BARRIER FUNCTIONS

A basic safety question is whether a controller can keep a system within its constraints indefinitely. Viability theory studies the initial conditions from which such evolutions exist (Aubin et al., 2011). Hamilton–Jacobi (HJ) reachability characterizes reachable sets through dynamic games (Mitchell et al., 2005; Bansal et al., 2017). Discounted safety Bellman equations connect this analysis to reinforcement learning (Fisac et al., 2019), and deep reachability methods approximate value functions in high-dimensional systems (Bansal & Tomlin, 2021). Related formulations learn infinite-horizon reach-avoid strategies (Li et al., 2022); Ganai et al. (2024) survey the broader connection between HJ analysis and RL.

Control barrier functions (CBFs) express invariance through inequalities on admissible controls (Ames et al., 2016; 2019), including discrete-time formulations (Agrawal & Sreenath, 2017) and their use within model predictive control (Zeng et al., 2021). Control barrier-value functions connect reachability values to barrier-based filtering (Choi et al., 2021); HJ refinement can improve candidate barriers (Tonkens & Herbert, 2022). Learning also supplies barriers from expert demonstrations under dynamics and regularity assumptions (Robey et al., 2020). Policy neural CBFs instead evaluate a fixed policy's maximum-over-time cost (So et al., 2024b). The survey by Dawson et al. (2023) places these approaches within learning of safety and stability certificates.

WISH shares the use of a state-dependent value to select admissible controls. Its numerical levels measure minimum cumulative disturbance effort to failure under a specified fallback, rather than the worst future value of a supplied state-space margin. The fixed-policy qualification is important: this resistance depends on the fallback's ability to recover and is not automatically the value of an optimal controller–adversary game. Learning a useful certificate and verifying that an approximation satisfies its required inequalities remain separate problems.

D.2 Predictive filters, shielding, and recovery

Safety filters allow a task policy to operate while checking whether its actions preserve a safe continuation (Hsu et al., 2024). Predictive safety filters use constrained planning and a backup safety policy, including uncertainty in learned dynamics (Wabersich & Zeilinger, 2021). Model predictive shielding checks recoverability under a backup policy (Bastani, 2021), with statistical variants addressing stochastic dynamics (Bastani et al., 2021). Backup CBFs define an implicit invariant set by integrating a fixed backup policy forward (Chen et al., 2021). These approaches already connect fallback behavior to acceptance of a nominal action.

Other intervention mechanisms use different safety representations. Formal shields synthesize action restrictions or corrections from a model and temporal-logic specification (Alshiekh et al., 2018). A learned safety layer can correct continuous actions using a local constraint model (Dalal et al., 2018). Safety critics transfer learned safety information to new tasks (Srinivasan et al., 2020). Recovery RL separates task and recovery policies, using a learned critic to identify when the recovery policy should take over (Thananjeyan et al., 2021). Thus, neither a reusable safety critic nor recovery-based intervention is unique to WISH. The distinction is the critic's disturbance-effort interpretation and the corresponding admissibility condition that preserves a cumulative budget under the exact modeled dynamics.

D.3 Stochastic and risk-sensitive safety

Under stochastic dynamics, safety must distinguish the probability of an unsafe event from a favorable mean value. Stochastic barrier certificates bound the probability of reaching an unsafe set (Prajna et al., 2007). For discrete-time systems, Cosner et al. (2023) derive finite-horizon probability bounds using martingale arguments when a barrier condition holds in expectation, with additional hypotheses including an upper bound on the barrier. Cosner et al. (2024) use Freedman's inequality and bounds on martingale differences and conditional variance to obtain stochastic safety bounds without requiring that upper bound on the barrier itself. Stochastic neural CBFs also investigate learned certificates for stochastic systems (Zhang et al., 2025a). Risk-constrained RL provides another formulation, optimizing policies subject to chance or conditional-value-at-risk constraints on cumulative costs (Chow et al., 2018).

The expectation-based counterexample in the main text concerns the particular action-admissibility condition evaluated there: positive outcomes can compensate for failure outcomes, allowing an action with substantial one-step failure probability. It does not imply that expectation-based barriers lack probabilistic guarantees; Appendix B.10 gives such a bound for the baseline. WISH takes a different route: under its exact-value and initial-state conditions, failure under the filter requires accumulated disturbance cost to reach the specified budget. A probability bound then follows from the distribution of that accumulated cost. These routes require different assumptions, and neither is claimed to dominate the other for every system or disturbance distribution.

D.4 Latent safety and partial observability

When safety is recognizable from observations but difficult to express in physical coordinates, latent models provide a space in which to predict the consequences of actions. Latent Safety Filters use a learned failure classifier and world-model dynamics for latent HJ reachability (Nakamura et al., 2025). LatentCBF identifies classifier-margin saturation and inaccurate nominal-action values as obstacles to smooth filtering, and addresses them using gradient penalties and training data from both nominal and safety policies (Nakamura et al., 2026). Both already learn safety from failure labels. WISH builds on this setting while asking a separate question: how should nominally safe states be ranked by their resistance to failure?

Representation and uncertainty raise additional challenges. UNISafe uses calibrated epistemic uncertainty to address out-of-distribution failures (Seo et al., 2025). AnySafe parameterizes constraints with images so that a latent filter can adapt its safety specification at runtime (Agrawal et al., 2026). Kim et al. (2026) investigate whether world-model representations retain the information needed to understand partially observable safety constraints. Adjacent approaches account for uncertainty in learned perception through measurement-robust CBFs (Dean et al., 2021), learn robust output-feedback barriers from demonstrations (Lindemann et al., 2024), or use partial-model shielding under partial observability (Carr et al., 2023).

Changing WISH's disturbance budget adjusts required resistance for a fixed failure specification. It does not recover missing safety information, calibrate epistemic uncertainty, or certify a learned observation model. Robustness to the chosen imagination-noise channel therefore differs from robustness to every possible perception or model error; physical safety requires the additional consistency conditions discussed in the transfer appendix.

D.5 Safe RL and world-model learning

Safe learning encompasses both restrictions during exploration and requirements on the policy eventually deployed (Brunke et al., 2022). Constrained Policy Optimization learns policies subject to expected cumulative cost constraints (Achiam et al., 2017). SafeMDP explores safely reachable states using Gaussian-process confidence information (Turchetta et al., 2016), while RL-CBF combines learned dynamics uncertainty with barrier-based intervention during policy learning (Cheng et al., 2019). SABRE learns with binary state-action safety feedback from an offline oracle (Bennett et al., 2023). That oracle supplies action-safety information; it should not be conflated with WISH's labels identifying whether a state is already in failure. These approaches reinforce that binary supervision alone is not a new safety-learning capability.

World models support prediction and policy learning from compact representations (Ha & Schmidhuber, 2018a; Hafner et al., 2019; 2023). LAMBDA uses Bayesian world models for constrained policy optimization, combining optimistic task estimates with pessimistic constraint estimates (As et al., 2022). More recent diffusion-world-model work evaluates candidate trajectories and refines actions with Langevin dynamics to balance reward and safety (Liu et al., 2026). Thus, learning safety-aware behavior in imagined trajectories also predates WISH. WISH uses a frozen world model to evaluate and improve a fallback and to construct a filter for nominal task actions. Expected constraint costs, failure probabilities, and minimum attack effort need not induce the same state ranking, and safe imagined rollouts do not by themselves establish safe data collection in the physical environment.

D.6 Adversarial and robust control

Robustness can be modeled through uncertainty in transition probabilities, disturbances acting on a fixed system, or uncertainty in the model itself. Robust dynamic programming optimizes against sets of transition laws (Iyengar, 2005; Nilim & El Ghaoui, 2005). Robust adversarial RL trains controllers against destabilizing disturbance policies (Pinto et al., 2017). ISAACS jointly learns a fallback and disturbance policy and uses robust forward rollouts for runtime certification (Hsu et al., 2023); related robust safe RL uses reachability games for constrained policy optimization (Li et al., 2023). RAMBO-RL instead adversarially optimizes learned dynamics during offline policy training (Rigter et al., 2022). These adversaries differ in their objectives and the quantities they can change.

Disturbance robustness also has a substantial control-theoretic foundation. Input-to-state safe CBFs relate disturbance magnitude to an enlargement of the invariant safe set (Kolathaya & Ames, 2019). Dissipation-based reachability uses storage functions to bound finite-horizon reachable sets under energy-bounded disturbances and integral quadratic constraints (Yin et al., 2020). Consequently, accounting for cumulative disturbance energy is not itself new. WISH evaluates the cheapest successful attack against a particular fallback through a specified channel in a frozen latent model. Its learned attack cost can overestimate resistance when the adversary misses a cheaper failure; robust optimization of model uncertainty and verification of the learned value address distinct issues.

D.7 FALSIFICATION AND MINIMUM-EFFORT FAILURE

Failure search asks how a system can violate a specification, but the search objective determines what the resulting vulnerability means. Black-box validation includes falsification, most-likely failure search, and failure-probability estimation (Corso et al., 2021). S-TaLiRo searches for temporal-logic violations using quantitative robustness (Annapureddy et al., 2011); Breach supports simulation-based verification and parameter synthesis (Donzé, 2010). Adaptive stress testing uses RL to search for likely failure trajectories (Lee et al., 2020), with extensions transferring failures from low- to high-fidelity simulation (Koren et al., 2021). Minimum-tolerance falsification searches for small parameter deviations that cause specification violations (Zhang et al., 2024).

Minimum-action methods provide an adjacent connection: they minimize a large-deviation action to find rare transition paths in small-noise systems (E et al., 2004). This connection motivates care in interpreting disturbance effort as likelihood. WISH uses $-\log q(\epsilon)+\log q(0)$ as its per-step cost. For a fixed horizon and independent noises, the shift is constant across disturbance sequences, so minimizing the summed cost is equivalent to maximizing their joint density. For variable failure times, that shift depends on trajectory length, and density optimization is not itself a failure-probability calculation. The finite-horizon budget bound is a separate statement, not a small-noise asymptotic approximation.

LEAP is the direct predecessor: it defines minimum cumulative disturbance effort to failure under a fixed fallback, derives continuous-time barrier properties and budget-based filtering, and learns potentials and improved fallbacks through alternating RL (So et al., 2026). WISH inherits that construction rather than introducing minimum-effort safety values or alternation itself. Its extension develops discrete-time latent dynamics, filtered-policy probability analysis for modeled stochasticity, and practical learning and deployment in world-model imagination. The comparison with falsification concerns the state-dependent value's use for recovery and online filtering, rather than a claim that failure search has never informed controller design.