

World Model Robustness via Surprise Recognition

Geigh Zollicoffer*, Tanush Chopra*, Mingkuan Yan, Xiaoxu Ma, Kenneth Eaton, Mark Riedl
Georgia Institute of Technology

{gzollicoffer3, tchopra32, myan71, xma394, keaton30, riedl}@gatech.edu*

Abstract

AI systems deployed in the real world must contend with distractions and out-of-distribution (OOD) noise that can destabilize their policies and lead to unsafe behavior. While robust training can reduce sensitivity to some forms of noise, it is infeasible to anticipate all possible OOD conditions. To mitigate this issue, we develop an algorithm that leverages a world model’s inherent measure of surprise to reduce the impact of noise in world model-based reinforcement learning agents. We introduce both multi-representation and single-representation rejection sampling, enabling robustness to settings with multiple faulty sensors or a single faulty sensor. While the introduction of noise typically degrades agent performance, we show that our techniques preserve performance relative to baselines under varying types and levels of noise across multiple environments within self-driving simulation domains (CARLA and Safety Gymnasium). Furthermore, we demonstrate that our methods enhance the stability of two state-of-the-art world models with markedly different underlying architectures: Cosmos and DreamerV3. Together, these results highlight the robustness of our approach across world modeling domains. We release our code at <https://github.com/Bluefin-Tuna/WISER>.

1. Introduction

AI systems that operate in the real world can often encounter noise, distractions, environmental interference, transmission errors, or new semantic classes of objects. When these occur, AI policies—the part of the AI system that translates from observation to action—can become unpredictable, unreliable, or unsafe. When observations contain out-of-distribution noise, the policy may respond inappropriately to the noise, resulting in execution that is unsafe to the agent or to any human operators in the vicinity.

At the heart of the challenge, an AI system cannot know what is noise and what is actionable information in the ob-

servation. While AI system developers may do their best to train systems that are robust to as many types of noise as possible, in real-world settings, it is impossible to anticipate all possible ways in which sensor information may be corrupted. That is: AI systems can always encounter noise and distractors that are interpreted incorrectly. Yet, we may be able to design AI systems that degrade gracefully by being able to (a) identify when parts of their observations are likely to be noise versus actionable information and (b) learn how to make smarter decisions despite the lack of actionable information.

In the context of world model-based deep reinforcement learning agents, we present a technique for improving the resilience of an agent when *parts* of an observation are likely to be noise and/or completely corrupted. We use the agent’s world model to evaluate the amount of *surprise* in the observational data. We apply this to multi-sensor and single-sensor agent settings. Multi-sensor settings are those in which an agent has multiple sensors to which a policy model responds, or has multiple representations of the same raw observational data. For example, in automated driving, a policy may integrate inputs from cameras, LiDAR, and radar, or fuse both pixel-level images and semantic maps derived from the same visual stream. It is likely that the underlying cause of the noise does not affect all sensors or representations equally. In this setting, the surprise measure is used to help identify which sensors are likely to distract the agent and to disable those sensors.

While the above is useful for agents that have multiple sensors, such as self-driving cars, some agents may only have a single sensor. In the single-sensor setting, we use the world model’s abnormal reconstruction of the observation as a signal to help avoid making decisions on observations that are likely to be distractions or misinterpreted by the agent. In both settings, we show that our technique is able to reduce the amount of noise introduced to the world model’s internal state.

Although the introduction or removal of noise typically degrades agent performance, we show that our techniques preserve performance relative to baselines under varying types and levels of noise across multiple environments

*Equal contribution

within self-driving simulation domains (CARLA [6] and Safety Gymnasium [18]). Furthermore, we demonstrate that our methods enhance the stability of two state-of-the-art world models with markedly different underlying architectures: Cosmos [24] and DreamerV3 [13]. Together, these results demonstrate the robustness of our approach across world modeling domains.

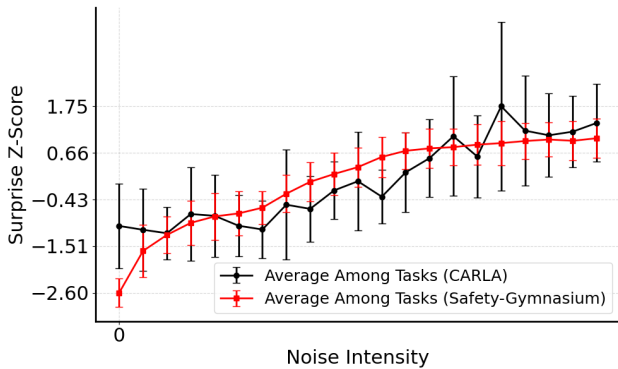


Figure 1. The change in the world model’s measure of surprise as noise increases in all tested CARLA [6] and Safety-Gymnasium [6] environments over 15000 sampled steps per intensity. We leverage this insight to understand the degree to which surprise signal can help identify noise given the absence of ground truth.

2. Related Work

Improving the robustness of reinforcement learning (RL) agents against noise capabilities largely remains a difficult challenge [30]. Procedural generation has been employed to improve agent robustness to novel scenarios [5], often relying on data augmentation techniques [31]. However, these approaches can lead to high sample complexity [22]. A further complication is the burden on the policy to learn a wide range of representations for rapid adaptation [7].

Recent work has also focused on identifying irrelevant components of the observation space to enhance generalization [3, 16, 28, 34]. Such methods often rely on foundation models or Siamese-like architectures during the training process, which can significantly increase training time [9]. Moreover, these frameworks may require an augmented dataset for initial training on similar OOD situations to the target domain, or are tested in cases where the task-relevant information is always present. Simple Gaussian noise, despite the OOD situation, has been proposed to help models learn OOD scenarios, but might not be effective in certain situations [25].

A promising approach for improving model robustness is the use of dropouts across various sensors or multimodal representations [21, 26, 27, 32]. However, these methods typically ignore the issue of determining which sen-

sors to mask out or assume that the agent can already identify which sensors have failed or have become corrupted—a challenging problem in itself [20]. Another limitation is that such dropout techniques often provide limited insight into how accurately an agent can predict actions using only the remaining modalities, and often leave the dropout strategy unchanged in the evaluation domain [27].

World models [11] have recently been shown to have some success in out-of-distribution scenarios [36] that cannot be taken into account during training. Specifically, results have shown that utilizing the world model’s measure of surprise does appear to be a useful heuristic in detecting if an agent’s observation is out of distribution. We investigate whether world models can mitigate the issues seen in previous dropout-related frameworks in RL from the perspective of state representation for novel situations.

Fault Detection, Isolation, Recovery (FDIR) [2, 29] is a common technique for removing noise, which, at a high level, aims to shut off or mitigate the impact of known faulty sensors. For physical systems, such as automotive applications, this typically manifests through the deployment of multiple redundant sensors that measure the same physical state. This redundancy is typically used to enable potential cross-validation among sensors, allowing faulty readings to be detected and suppressed while preserving overall system integrity [4].

In our work, we do not assume access to redundant sensors, as they would experience the same OOD noisy failures. Instead, we assume the potential of having different data per sensor or a single sensor with different representations of the data. We design an FDIR algorithm centered on world models so the agent can focus on consistent and informative inputs when noise or sensor failures are observed.

3. Preliminaries

Partially observable Markov Decision Processes We study sequential Partially Observable Markov decision processes (POMDPs) denoted by the tuple $M = (\mathcal{S}, \mathcal{A}, \mathcal{T}, r, \Omega, O)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{T}$ is the transition distribution $\mathcal{T}(s_t | s_{t-1}, a_{t-1})$, r is the reward function, Ω is an emissions model from ground truth states to sensor information, and $O(s', a)$ is the sensory information distribution which emits an observation $x_* \sim O(s', a)$ at each step [1].

DreamerV3 World Model We conduct experiments using our methods primarily within the DreamerV3 framework [13], which is based on a Recurrent State-Space Model (RSSM). A typical RSSM incorporates both a Variational Autoencoder (VAE) and a Recurrent Neural Network (RNN) [10, 11]. DreamerV3 first learns a world model and then uses it to simulate rollouts for training a policy model. Note that:

- x_* is the set of sensory information emitted from the state.
- x_t is the *encoded* set of sensory information emitted from the state.
- h_t is the encoded history/hidden state of the agent.
- z_t is an encoding of the current sensory information x_t and h_t that incorporates the learned dynamics of the world.
- $s_t = (z_t, h_t)$ is the agent’s compact model state.
- At each step, the agent takes an action based on the compact model state (i.e. $\pi(s_t) = \pi((z_t, h_t))$)

For each representation of state s_t , DreamerV3 defines six additional learned transition distributions—each conditionally independent—based on the trained world model:

$$\text{DreamerV3: } \begin{cases} \text{Sequence model: } h_t = f_\phi(h_{t-1}, z_{t-1}, a_{t-1}) \\ \text{Representation model: } q_\phi(z_t | h_t, x_t) \\ \text{Dynamics predictor: } p_\phi(\hat{z}_t | h_t) \\ \text{Image prediction model: } p_\phi(\hat{x}_t | h_t, z_t) \\ \text{Reward prediction model: } p_\phi(\hat{r}_t | h_t, z_t) \\ \text{Continue prediction model: } p_\phi(\hat{c}_t | h_t, z_t) \end{cases} \quad (1)$$

where ϕ describes the parameter vector for all distributions optimized. The loss function during training [13] is:

$$\mathcal{L}(\phi) \doteq E_{q_\phi} \left[\sum_{t=1}^T (\beta_{\text{pred}} \mathcal{L}_{\text{pred}}(\phi) + \beta_{\text{dyn}} \mathcal{L}_{\text{dyn}}(\phi) + \beta_{\text{rep}} \mathcal{L}_{\text{rep}}(\phi)) \right].$$

where:

$$\begin{aligned} \mathcal{L}_{\text{pred}}(\phi) &\doteq -\ln p_\phi(x_t | z_t, h_t) \\ &\quad -\ln p_\phi(r_t | z_t, h_t) - \ln p_\phi(c_t | z_t, h_t) \\ \mathcal{L}_{\text{dyn}}(\phi) &\doteq \max \left(1, KL[sg(q_\phi(z_t | h_t, x_t)) \| p_\phi(z_t | h_t)] \right) \\ \mathcal{L}_{\text{rep}}(\phi) &\doteq \max \left(1, KL[q_\phi(z_t | h_t, x_t) \| sg(p_\phi(z_t | h_t))] \right) \end{aligned}$$

and $sg(\cdot)$ is a stop gradient operator.

4. Multiple Sensor Representation Selection

In this section, we address the multiple-sensor representation setting and describe how a surprise measure on the world model identifies and eliminates sensor representations that are likely to distract the agent with noisy data. The multiple sensor setting could mean that an agent has multiple sensors that collect different data, such as in the case of an autonomous vehicle with cameras with different views. It could also mean an agent with a single sensor but representing that sensor data in different ways.

We assume a *world model*, such as that used in DreamerV3 [13]. The world model learns the world transition dynamics and is used to more efficiently train the policy model by predicting state-action-state transitions, effectively separating the agent from the true environment. Although many

of the world model’s learned predictors are rarely used at inference time, we find that the model’s surprise grows proportionally with the intensity of injected noise. This relationship is shown in Figure 1 for two domains. The trained world model’s measure of Bayesian surprise [17] is the divergence between the world model’s learned posterior and prior of the next state:

$$KL[q_\phi(z_t | h_t, x_t) \| p_\phi(z_t | h_t)] \quad (2)$$

This measure naturally provides a correlated signal for detecting unexpected noise. When the model consistently registers unusually high surprise without corresponding structure in the environment, it may indicate that the observations are dominated by noise rather than meaningful dynamics. Building on this hypothesis, we introduce a surprise-guided rejection sampling method that mitigates corruption in the world model’s predicted latent state while maintaining an $\mathcal{O}(n \log n)$ computational cost. during inference time.

Surprise-Guided Multi-Representation Rejection Sampling for Confident Decision Making

We investigate the world model’s ability to *adaptively switch representations* in an online failure setting, where sensor failures occur without prior knowledge. In the multi-sensor setting, the agent samples a set of M high-dimensional sensory observations at each time step,

$$\mathbf{x}_* = \{y_t^{(1)}, y_t^{(2)}, \dots, y_t^{(M)}\},$$

from which it forms a fused observation

$$x_t = f_{\text{enc}}(\mathbf{x}_*),$$

where f_{enc} denotes the encoder or sensor-fusion module. x_t is then used to sample the latent state z_t via the representation model (Eq. 1).

During evaluation, if any subset of sensory inputs $y_t \in \mathbf{x}_*$ exhibits abnormally high surprise, indicating a potential corruption or failure, the model avoids relying on the full fused observation x_t . Instead, it samples a latent state prediction $z_t^{(i)}$ conditioned on a fused observation $x_t^{(i)}$ formed by a subset of uncorrupted inputs $\mathbf{x}_*^{(i)} \subseteq \mathbf{x}_*$. This selection is formulated as an optimization that chooses the input subset minimizing the Bayesian surprise between the representation model and the dynamics predictor (Eq 1):

$$\arg \min_{\mathbf{x}_*^{(i)} \subseteq \mathbf{x}_*} KL[q_\phi(z_t^{(i)} | h_t, x_t^{(i)}) \| p_\phi(z_t | h_t)]. \quad (3)$$

The resulting sampled latent $z_t^{(i)}$ serves as an *alternative latent state representation*, which is then used to select an action $a_t \sim \pi((z_t^{(i)}, h_t))$.¹

¹Optionally, the optimization can be relaxed to respect a set of required sensors (see Algorithm 2).

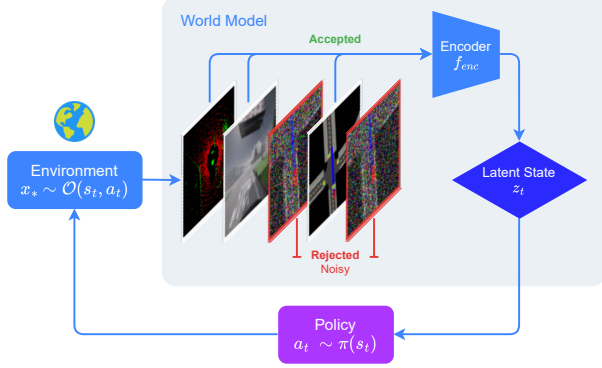


Figure 2. Multi-Sensor Rejection Sampling. In practice, we utilize Algorithm 2 to simulate this behavior in the agent.

Multi-Representation Dropout Training To simulate the removal of sensory inputs, we apply a mask-based sensor dropout. Naturally, the absence of expected sensory inputs induces increased surprise within the world model. Therefore, to prepare the agent for dropout variation in its sensors, we expose it to projections onto random subspaces of the full representation space via random dropout of representations—a form of state representation learning [7]. As shown in Algorithm 1, this encourages the world model to learn a latent distribution that remains robust even when some components of the sensory inputs are missing. By experiencing these partial views during training, we observe (Figure 16 Appendix E.3) that the world model is still capable of inferring consistent and informative latent states despite variability in the available representations.

5. Multiple Sensor Experiments

For our main results, we primarily focus on settings that naturally involve multiple sensors, such as automated driving tasks. Unless stated otherwise, we use the default sensors and reward functions provided by each environment for each respective task.

Environments We show the generalization of our proposed method through experiments conducted primarily in the following domains:

- **Safety Gymnasium** [18] is a benchmark suite designed to evaluate the trade-off between performance and safety in reinforcement learning, extending the traditional RL paradigm by introducing environments where agents must complete goal-oriented tasks while simultaneously minimizing safety violations such as collisions or entering hazardous regions. The environments include a variety of robotic control and navigation tasks that simulate real-world constraints, providing explicit cost signals associated with unsafe behaviors. Our experiments use the default camera sensors given by the environment (see Ap-

pendix A for details). Safety Gymnasium and the SafeDreamer [15] baseline assume that the training and evaluation environments are identical. In practice, this assumption may not hold, and our experiments test this by injecting out-of-distribution data..

- **CARLA** [6] is an open-source urban driving simulator. It was developed as a free platform for autonomous driving research. Alongside the core simulator code and APIs, CARLA includes a library of custom-designed urban environments and digital assets (e.g., buildings, vehicles, pedestrians) that are freely available for use. The simulator supports flexible configuration of sensor suites, such as RGB cameras, Bird Eye view, LiDAR, GPS, and collision detectors.

We list each utilized sensor in detail in Appendix A. In each environment, we simulate sensor failures by injecting various types of structured and unstructured out-of-distribution noise into the sensor data. Examples include gaussian, occlusion, glare, jitter, chromatic aberration (chrome), and latency. We provide example visualizations of noises applied to the CARLA Bird Eye View (BEV) sensor in 7. Additionally, all noises are described in Appendix B.2. The agent is never notified of any properties of the failure scenario. In all experiments, there is always at least one sensor that is unaffected.

Agent Models In this setting, we focus our experiments on the State-of-the-Art DreamerV3 world model [13]. For each experiment, the *Base* agent is a model trained normally. An *Augmented* agent is the base model trained on a dataset augmented with Gaussian noise. The *Random Masked Encoder* (RME) [27], is a technique that also employs masked based sensor dropout, however does not provide a means to decide which sensors should be masked out. For the multi-sensor domain, our rejection sampling technique is labeled as *Confident Representation* (as to not be confused with single-sensor results), which consists of agents equipped with the surprise mechanism, dropout training, and multi-sensor selection technique described in Section 4. For all experiments during inference time, we deploy Algorithm 2, as our proposed multi-sensor selection algorithm, designed to run in $O(n \log n)$ without parallel optimizations. Unless otherwise specified, each experiment data point represents an accumulation of 15,000 steps across multiple independent and identically distributed environments for a particular method within the given task and noise scenario. We evaluate each method in a fully closed-loop setting, where the agent’s actions directly influence subsequent observations. For both Safety Gymnasium and CARLA, we focus on the agent’s reward in all tasks. In addition, for the Safety Gymnasium domain, we also focus on cost, an empirical safety score that the Safety Gymnasium domain provides.

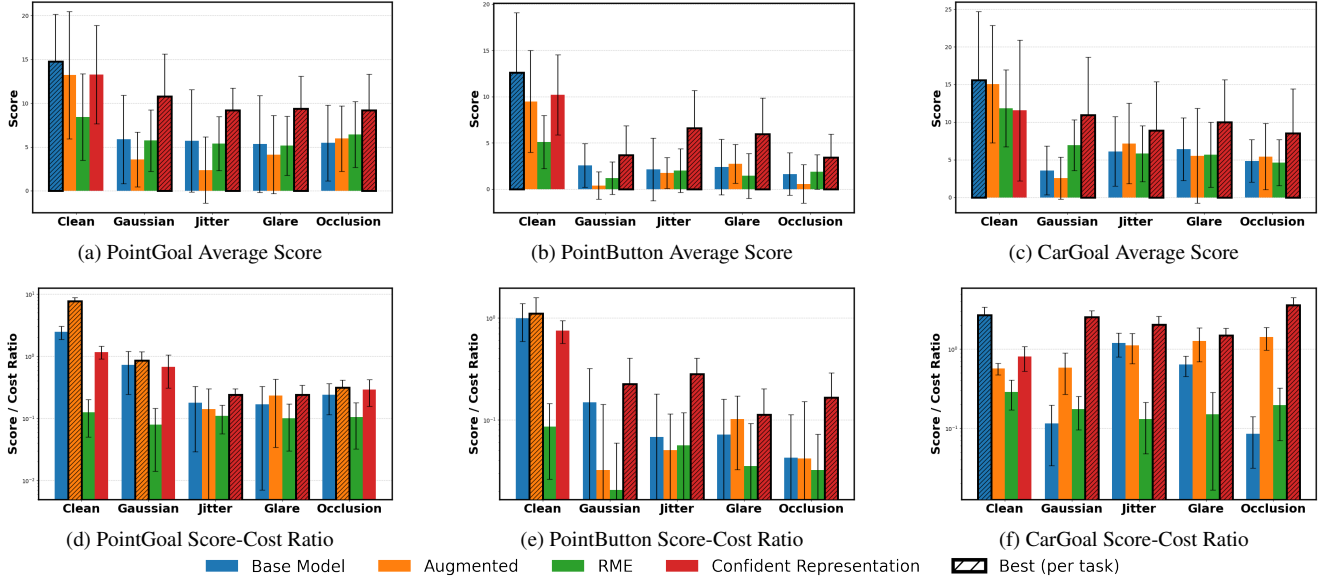


Figure 3. Performance (top) and Score-Cost ratio (bottom) across three Safety Gymnasium tasks. We display the Score-Cost ratio to measure how safely the task can be accomplished with respect to the score. Each column corresponds to a noise type, with clean being the nominal setting.

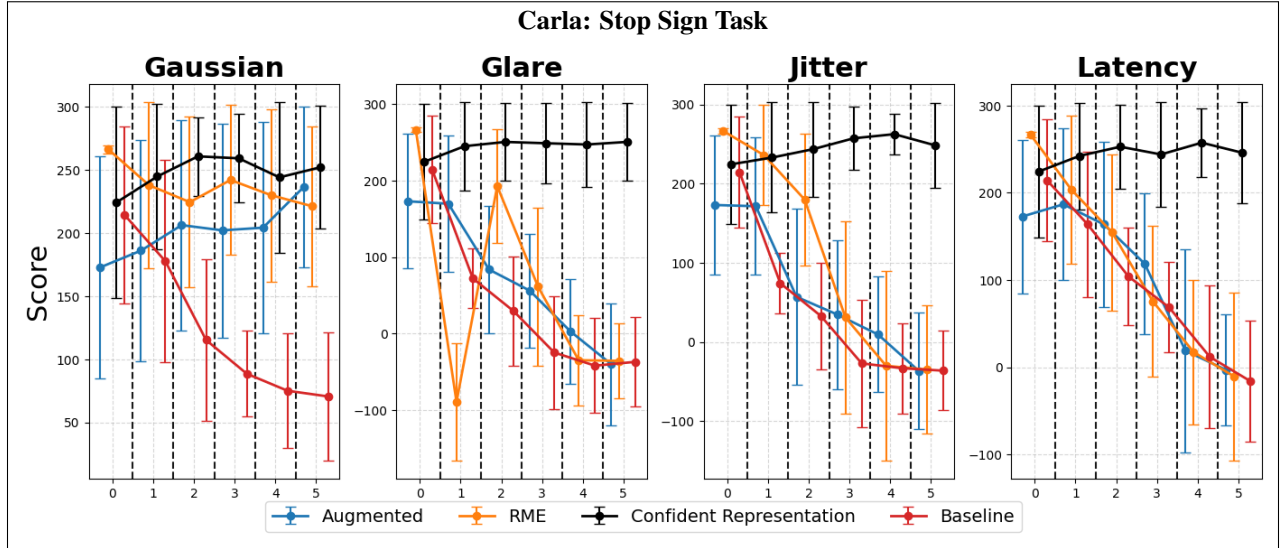


Figure 4. Agent performance as the number of sensor failures (from 0-5 sensors) during the stop sign task. We observe that as the sensors begin to become affected, we are able to reduce the effect that noise has on the main policy. For all 72 tested settings over the Stop Sign, Right Turn, and Four Lane Driving Tasks, see Appendix 10.

Results Figure 3 illustrates our results on the Safety Gymnasium environment. Our Confident Representation method is capable of consistently achieving the highest level of robustness on a variety of simulated sensor failures. It is also able to accomplish said tasks with a competitive score to cost ratio compared to other methods tested in failure scenarios.

In CARLA, we can have up to 5 sensor failures. Results

are shown in Figure 4. We observe that our Confident Representation method is capable of identifying failed sensors to shut off during inference time. In comparison to other methods, as sensors begin to fail, the agent is able to identify and use representations that provide meaningful signals and achieve high rewards.

Also, despite not being trained on the exact Gaussian noise that the Gaussian failure was trained on, Confident

Representation selection is capable of mirroring the performance of the Augmented baseline in Figure 4. Finally, we compare our $O(n \log n)$ sensor selection technique (Algorithm 2 in Appendix E.4) to an exhaustive search of sensors and show in Figure 9 that our technique can expect similar performance despite being magnitudes lower in computational complexity.

6. Single Sensor Representation Selection

We next consider a setting where the agent samples a single high-dimensional observation $x_* \in \mathcal{X}$ at each decision step from the environment, which may vary from a clean, informative signal to a heavily corrupted input with limited recoverable information. As with the previous multi-sensor setting, the OOD noise resulting from potential sensor failure may cause the agent to take the wrong action, or may have no effect on action choice.

To design an algorithm capable of improving robustness, we first categorize the possible forms of x_* into three cases:

1. $x_* = x_t$: The observation is clean, and the agent can directly encode and rely on it for decision-making.
2. $x_* = x_{light}$: The observation contains some noise, but the underlying informative signal can still be recovered or emphasized through selective processing or denoising.
3. $x_* = x_{heavy}$: The observation is dominated by noise to the extent that the true signal is no longer recoverable and should therefore be ignored.

Given the possible classes of x_* , we enable the agent to switch between two modalities during inference: a *predictive mode*, wherein the agent rejects x_{heavy} and the agent instead relies on its internally predicted next state from the world model, and a *ground-truth mode*, wherein the agent accepts the sampled observation x_t or its denoised version $x_{light} = D(x_*)$, for some choice of denoiser $D(\cdot)$. In essence, when the agent determines that the input data is sufficiently corrupted such that it jeopardizes action selection, it prefers its internal predictions of the world dynamics over what it observes.

To decide if a observation x_* should be rejected, a rejection score is assigned by a function, $M(x_*)$, and is rejected if it reaches a threshold τ . For our experiments we employ $M(\cdot)$ as an expected reconstruction loss of the unconditioned ($h_t = h_0$) observations: $\frac{1}{|N|} \sum_{i \in N} |x_t^{(i)} - \hat{x}_t^{(i)}|$ where $\hat{x}_t = \mathbb{E}_{p_\phi(x_t|z_t, h_0)}[x_t]$, and employ $D(\cdot)$ to be the predicted posterior representation generated by the world model variational autoencoder: $\mathbb{E}_{p_\phi(x_t|z_t, h_t)}[x_t]$, as proof of concept to identify abnormal sensory input. Intuitively, $M(\cdot)$ and τ are defined to indicate the presence of a failed sensor, signaling that directly processing x_* would likely lead to unpredictable behavior. The objective of our rejection sampling mechanism is to enable the agent to produce

predictable and conservative responses in the presence of unknown sensor failures. Since the purpose of rejection sampling is to preserve the predicted latent state from corruption, downstream policies can then implement conservative strategies—such as pulling over—minimizing the risk of unpredictable behavior occurring during their execution.

Aligning Context During predictive mode, the agent relies on its own imagined or simulated dynamics, which can gradually drift from the true environment state. The agent transitions from predictive mode to ground-truth mode when an observation is accepted as being x_t or x_n . When the agent has been in predictive mode for a period, it becomes necessary to reconcile the incoming sensory input with the agent’s internal context or trajectory h_t . Upon re-entering ground-truth mode, the agent performs a *context reset*—realigning its latent representation or belief state to ensure consistency between the accepted observation x_* and its previously predicted trajectory. This step prevents discontinuities in state estimation and allows the agent to resume grounded interaction with the environment.

Conversely, when an observation is rejected after operating in ground-truth mode, the agent retains its most recent valid internal context as a persistent residual. This residual serves as a stabilizing prior, allowing the agent to maintain coherent belief evolution during predictive inference until a new, trustworthy observation is accepted. In effect, the agent alternates between correcting drift upon acceptance and preserving continuity upon rejection, ensuring robust operation across varying observation qualities. We illustrate this interleaving of switching between modes in Figure 5.

7. Single Sensor Experiments

We focus our experiments on two types of world models: a VAE-based model (DreamerV3) [13] and a diffusion-based model (Cosmos Predict-2.5) [24]. For unmodified components of the world models, we use the exact hyperparameters recommended in [13], [15], [24], and [8]. For all other additional modifications, we list exact hyperparameters used and hardware details in Appendix D.

In these experiments, the world model-based agent uses a single sensor and only receives a single representation at every time step. In the single representation setting, we consider the effect that different combinations of noise, intensity, and proportion have on an agent during the course of an episode. To empirically test the behavior of our proposed method, We empirically validate our results against the original world model (*Base*), Median Filtering (*Filter*) [31], and Hybrid RSSM (*HRSSM*) [28], which augments the RSSM framework to capture task-relevant dynamics while suppressing noise. We set τ to be 5 standard deviations from the average reconstruction loss during transitions in the normal environment. Figure 6 shows results

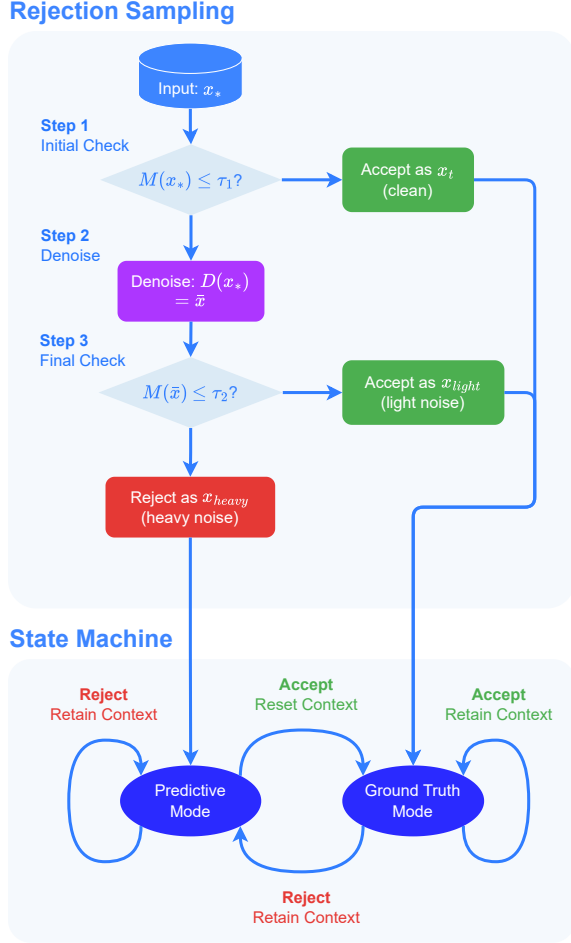


Figure 5. Rejection sampling process for noise classification and context state machine for the world model’s latent state.

from the Carla Four Lane task. Rejection Sampling consistently achieves the highest scores across all tested noise types, maintaining robustness as both the proportion and intensity of noise increase. Unlike other methods, it does not exhibit a decline toward lower negative scores under higher noise levels. We find that across all methods, the performance surfaces are generally skewed to the right, suggesting that increasing the proportion of corrupted observations has a stronger detrimental effect than increasing noise intensity.

7.1. Cosmos World Model

To further verify that the proposed rejection-based filtering mechanism generalizes beyond latent-state world models, we extend our study to the Cosmos Predict 2.5 world model—a Diffusion-based video world model that performs pixel-space prediction through large-scale diffusion and spatio-temporal self-attention.

We experiment with utilizing the *Rejection Sampling*

Table 1. Average PAI Quality score [35], between the base model and our proposed Rejection Sampling technique. Separated by noise type.

Aug Type	Base Model	Rejection Sampling	Avg Diff	Rel %
Chrome	0.774	0.808	0.034	3.13
Gaussian	0.767	0.810	0.043	5.00
Glare	0.726	0.809	0.083	11.98
Jitter	0.719	0.812	0.093	12.25
Occlusion	0.787	0.810	0.023	3.18
Overall	0.755	0.810	0.055	7.11

process displayed in Figure 5 in an alternative world model setting. We utilize each of the noises discussed in Appendix B.2 at a proportion of 75% to distort input videos, and monitor Cosmos-Predict 2.5’s generation quality. In this setting, we apply our proposed Rejection Sampling process on the Robot Pouring Task (See Appendix E.1 for prompt, visualization, and implementation details), and compare with the base Cosmos generation procedure.

For comparison with the standard Cosmos generation procedure, we monitor the PAI-Bench quality score [35]: An overall summary of eight scores pertaining to quality, consistency, and smoothness (Explicit descriptions can be found in Appendix E.1.2). We define $M(x_*)$ as the metric score measuring reconstruction quality:

$$M(x_*) = \frac{1}{CHW} \sum_{c,h,w} (x_*^{c,h,w} - \hat{x}_*^{c,h,w})^2$$

where $\hat{x}_* = f_\theta(x_*)$ is the immediate next frame generated by the model from conditioning only on the previous frame x_* , and C, H, W denotes the channel, height, and width dimensions. We evaluate each of the last N input frames x_* by computing their corresponding rejection score $M(x_*)$ (See Figure 11 for a visual aid). Similar to denoising methods found in unsupervised settings [19, 23, 37], we employ $D_\sigma(x_*)$ as a noise-and-denoise operator that: (1) encodes reference frame x_* to latent space, (2) corrupts the latent with Gaussian noise at strength σ , and (3) applies reverse diffusion denoising for reconstruction. The reverse diffusion process leverages the model’s learned Gaussian noise characteristics to remove the corruption and recover the true image. We report our findings in Table 1. We find that Rejection Sampling is capable of improving the relative PAI quality score by roughly 7% overall when the input video is corrupted. Additionally, we find that rejection sampling brings all augmentation types to a similar improved performance level (~ 0.81), regardless of their initial degradation severity.

8. Conclusion

We introduced a surprise-driven filtering framework that leverages the world model’s internal uncertainty to selec-

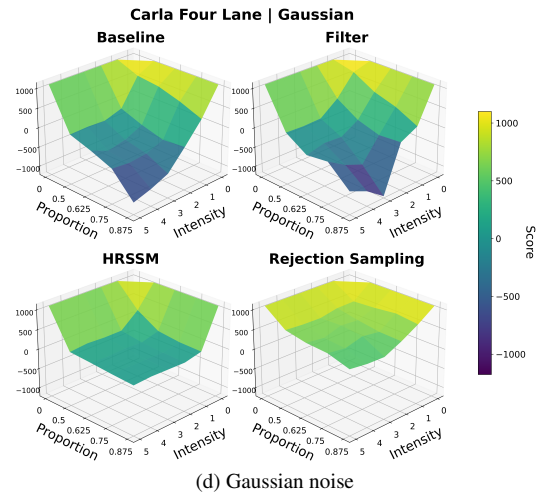
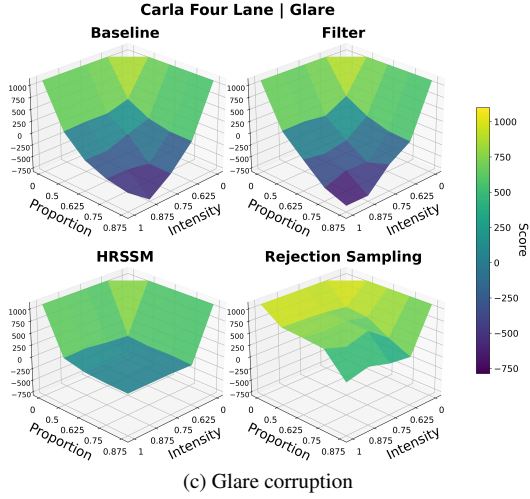
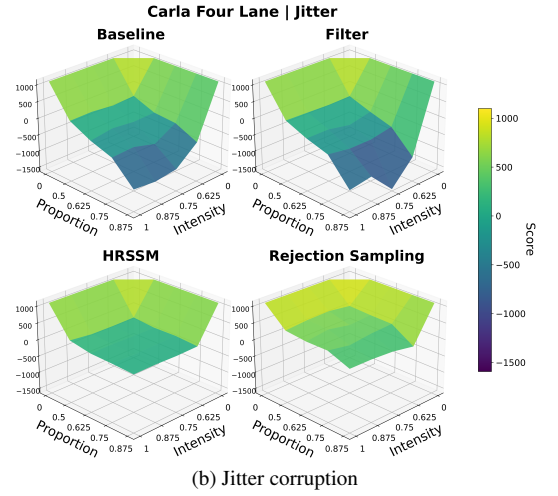
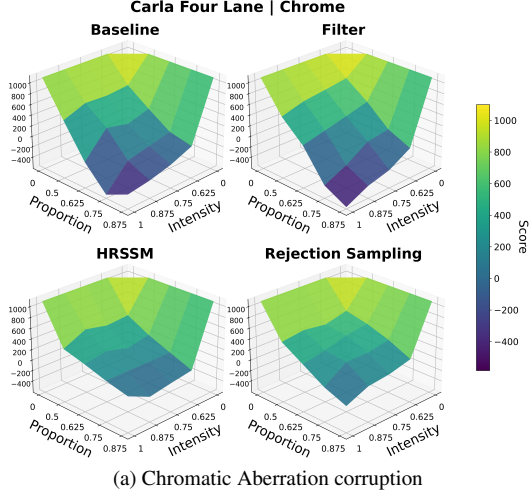


Figure 6. Visualization of policy behavior in the CARLA four-lane environment under different visual corruptions, accounting for 192 settings. Each subfigure shows the methods’ potential to safeguard the agent from a specific perturbation type. For results on other tasks see 8.

tively suppress unreliable sensory inputs. Our approach introduces two variants of rejection sampling—multi-sensor and single-representation—that enable world models to remain stable and effective under diverse noisy, out-of-distribution sensor failures. Empirical evaluations across self-driving tasks in CARLA and Safety Gymnasium, as well as across both diffusion-based and VAE-based world models, demonstrate that our method achieves robust performance without compromising adaptability. These results highlight the potential of surprise-based mechanisms for building world-model-based systems that maintain safety and stability in the face of unknown failures.

References

- [1] Karl Johan Åström. Optimal control of markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10:174–205, 1965. 2
- [2] Mogens Blanke, Michel Kinnaert, Jan Lunze, and Marcel Staroswiecki. *Diagnosis and Fault-Tolerant Control*. Springer, Berlin, Germany, 3rd edition, 2016. 2
- [3] Jen-Yen Chang, Thomas Westfechtel, Takayuki Osa, and Tatsuya Harada. Offline deep reinforcement learning for visual distractions via domain adversarial training. *Transactions on Machine Learning Research*, 2024. 2
- [4] Jie Chen and Ronald J. Patton. *Robust Model-Based Fault Diagnosis for Dynamic Systems*. Kluwer Academic Publishers, Boston, MA, 1999. 2
- [5] Karl Cobbe, Christopher Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. *CoRR*, abs/1912.01588, 2019. 2
- [6] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017. 2, 4, 11
- [7] Zeki Doruk Erden, Donia Gasmi, and Boi Faltings. Continual reinforcement learning via autoencoder-driven task and new environment recognition, 2025. 2
- [8] Dechen Gao, Shuangyu Cai, Hanchu Zhou, Hang Wang, Iman Soltani, and Junshan Zhang. Cardreamer: Open-source learning platform for world model based autonomous driving. *IEEE Internet of Things Journal*, pages 1–1, 2024. 6
- [9] Bram Grooten, Tristan Tomilin, Gautham Vasan, Matthew E Taylor, A Rupam Mahmood, Meng Fang, Mykola Pechenizkiy, and Decebal Constantin Mocanu. MaDi: Learning to Mask Distractions for Generalization in Visual Deep Reinforcement Learning. *The 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2024. URL: <https://arxiv.org/abs/2312.15339>. 2
- [10] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems 31*, pages 2451–2463. Curran Associates, Inc., 2018. <https://worldmodels.github.io>. 2
- [11] David Ha and Jürgen Schmidhuber. World models, 2018. Interactive version of the article at <https://worldmodels.github.io>. 2
- [12] Danijar Hafner. Benchmarking the spectrum of agent capabilities. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. Originally published as arXiv:2109.06780. 16
- [13] Danijar Hafner, Julian Pasukonis, Jimmy Ba, et al. Mastering diverse control tasks through world models. *Nature*, 640: 647–653, 2025. 2, 3, 4, 6, 15
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 18
- [15] Weidong Huang, Jiaming Ji, Borong Zhang, Chunhe Xia, and Yaodong Yang. Safedreamer: Safe reinforcement learning with world models. In *The Twelfth International Conference on Learning Representations*, 2024. 4, 6
- [16] Miles Hutson, Isaac Kauvar, and Nick Haber. Policy-shaped prediction: avoiding distractions in model-based reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 13124–13148. Curran Associates, Inc., 2024. 2
- [17] Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. In *Advances in Neural Information Processing Systems*. MIT Press, 2005. 3
- [18] Jiaming Ji, Borong Zhang, Jiayi Zhou, Xuehai Pan, Weidong Huang, Ruiyang Sun, Yiran Geng, Yifan Zhong, Josef Dai, and Yaodong Yang. Safety gymnasium: A unified safe reinforcement learning benchmark. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 2, 4, 11, 13
- [19] Minjong Lee and Dongwoo Kim. Robust evaluation of diffusion-based adversarial purification. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 134–144, 2023. 7
- [20] Daoliang Li, Ying Wang, Jinxing Wang, Cong Wang, and Yanqing Duan. Recent advances in sensor fault diagnosis: A review. *Sensors and Actuators A: Physical*, 309:111990, 2020. 2
- [21] Guan-Hong Liu, Avinash Siravuru, Sai Prabhakar, Manuela Veloso, and George Kantor. Learning end-to-end multimodal sensor policies for autonomous navigation. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 249–261. PMLR, 2017. 2
- [22] Robert Müller, Steffen Illium, Thomy Phan, Tom Haider, and Claudia Linnhoff-Popien. Towards anomaly detection in reinforcement learning. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, page 1799–1803, Richland, SC, 2022. International Foundation for Autonomous Agents and Multiagent Systems. 2
- [23] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification, 2022. 7
- [24] NVIDIA, :, Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, Prithvijit Chattopadhyay, Mike Chen, Yongxin Chen, Yu Chen, Shuai Cheng, Yin Cui, Jenna Diamond, Yifan Ding, Jiaojiao Fan, Linxi Fan, Liang Feng, Francesco Ferroni, Sanja Fidler, Xiao Fu, Ruiyuan Gao, Yunhao Ge, Jinwei Gu, Aryaman Gupta, Siddharth Gururani, Imad El Hanafi, Ali Hassani, Zekun Hao, Jacob Huffman, Joel Jang, Pooya Jannaty, Jan Kautz, Grace Lam, Xuan Li, Zhaoshuo Li, Maosheng Liao, Chen-Hsuan Lin, Tsung-Yi Lin, Yen-Chen Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Yifan Lu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Seungjun Nah, Yashraj Narang, Abhijeet Panaskar, Lindsey Pavao, Trung Pham, Morteza Ramezani, Fitsum Reda, Scott Reed, Xuanchi Ren, Haonan Shao, Yue Shen, Stella Shi, Shuran Song, Bartosz Stefaniak, Shangkun Sun, Shitao Tang, Sameena Tasmeen, Lyne Tchapmi, Wei-Cheng Tseng, Jibin Varghese, Andrew Z. Wang, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Jiashu Xu, Dinghao Yang, Xiaodong Yang, Haotian Ye, Seonghyeon Ye, Xiaohui Zeng, Jing Zhang, Qinsheng

- Zhang, Kaiwen Zheng, Andrew Zhu, and Yuke Zhu. World simulation with video foundation models for physical ai, 2025. 2, 6
- [25] Amartya Sanyal, Yaxi Hu, Yaodong Yu, Yian Ma, Yixin Wang, and Bernhard Schölkopf. Accuracy on the wrong line: On the pitfalls of noisy data for OOD generalisation. In *ICML 2024 Next Generation of AI Safety Workshop*, 2024. 2
 - [26] Younggyo Seo, Danijar Hafner, Hao Liu, Fangchen Liu, Stephen James, Kimin Lee, and Pieter Abbeel. Masked world models for visual control, 2023. 2
 - [27] Skand Skand, Bikram Pandit, Chanho Kim, Li Fuxin, and Stefan Lee. Simple masked training strategies yield control policies that are robust to sensor failure. In *8th Annual Conference on Robot Learning*, 2024. 2, 4
 - [28] Ruixiang Sun, Hongyu Zang, Xin Li, and Riashat Islam. Learning latent dynamic robust representations for world models. In *ICML*, 2024. 2, 6
 - [29] The MathWorks, Inc. What is fault detection, isolation, and recovery (fdir)? <https://www.mathworks.com/discovery/fdir.html>, 2025. Accessed: 2025-08-12. 2
 - [30] Frederik Träuble, Andrea Dittadi, Manuel Wuthrich, Felix Widmaier, Peter Vincent Gehler, Ole Winther, Francesco Locatello, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. The role of pretrained representations for the OOD generalization of RL agents. In *International Conference on Learning Representations*, 2022. 2
 - [31] Zaitian Wang, Pengfei Wang, Kunpeng Liu, Pengyang Wang, Yanjie Fu, Chang-Tien Lu, Charu C. Aggarwal, Jian Pei, and Yuanchun Zhou. A comprehensive survey on data augmentation, 2024. 2, 6
 - [32] Tao Yu, Zhizheng Zhang, Cuiling Lan, Yan Lu, and Zhibo Chen. Mask-based latent reconstruction for reinforcement learning. In *Advances in Neural Information Processing Systems*, 2022. 2
 - [33] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022. 18
 - [34] Amy Zhang, Rowan Thomas McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2021. 2
 - [35] Fengzhe Zhou, Jiannan Huang, Jialuo Li, and Humphrey Shi. Physical ai bench: A comprehensive benchmark for physical ai generation and understanding, 2025. 7, 16
 - [36] Geigh Zollicoffer, Kenneth Eaton, Jonathan C Balloch, Julia Kim, Wei Zhou, Robert Wright, and Mark Riedl. Novelty detection in reinforcement learning with world models. In *Forty-second International Conference on Machine Learning*, 2025. 2
 - [37] Geigh Zollicoffer, Minh N. Vu, Ben Nebgen, Juan Castorena, Boian Alexandrov, and Manish Bhattarai. Loric: Low-rank iterative diffusion for adversarial purification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(21): 23081–23089, 2025. 7

A. Sensors

CARLA [6]

- *Lidar*: Captures a 3D point cloud of the surrounding environment.
- *Camera*: Provides a forward-facing RGB view from the vehicle.
- *Bird’s-Eye View (BEV) Raw*: Displays the roadmap and surrounding vehicles from a top-down perspective.
- *Collision*: Detects collisions with objects in the environment.
- *Bird’s-Eye View with Traffic Signals*: Renders traffic lights and signs directly on the BEV.
- *Bird’s-Eye View with Waypoints*: Renders only waypoints on the BEV (GPS setting).
- *Bird’s-Eye View Full*: Renders all available semantic information on the BEV.

Safety Gymnasium [18]

- *Bird Eye View*: A view of the agent and the entire map from above.
- *Front Vehicle View*: The immediate front view of the agent.
- *Dash Camera View*: The immediate front view of the agent from the dash cam perspective.

B. CARLA Noise Corruption and Sensor Degradation

B.1. Environment and Representations

In CARLA [6], an open-source urban driving simulator, the agent is equipped with a diverse sensor suite comprising seven modalities (see A). We evaluate performance on three representative scenarios—Stop Sign, Right Turn, and Four Lane Driving—which test the agent’s ability to perceive, plan, and act under different traffic configurations.

B.2. Robustness under Visual Corruptions

We evaluate the robustness of the driving agent in the CARLA simulator by subjecting input images to five types of visual corruptions:

- *Chromatic Aberration*: introduces small spatial shifts between RGB channels, producing color fringing and edge distortions similar to lens dispersion.
- *Gaussian Noise*: adds pixel-level random variations, modeling low-light sensor interference.
- *Glare*: causes extreme overexposure that whitens the whole frame, erasing most visual information.
- *Jitter*: applies random shifts in contrast or brightness, simulating unstable or high-noise sensor conditions.
- *Occlusion*: masks random regions of the frame, simulating partial blockage of the camera view.

The goal is to test how well the agent can still perform its task under degraded visual inputs. Quantitatively, we plot 3D surface graphs of the task performance metric as a function of corruption intensity and proportion. Each surface corresponds to one CARLA task—Stop Sign, Right Turn, and Four-Lane Driving—and compares four methods: the baseline world-model agent, Median Filtering, our proposed Rejection Sampling, and HRSSM. For all tasks shown in Figure 8, our Rejection Sampling maintains a much smoother and higher performance surface, indicating stable behavior even under severe sensor degradation. This trend demonstrates that the proposed Filter not only delays the onset of failure but also preserves task continuity across extreme perturbations, achieving consistently superior robustness and reliability in all CARLA environments.

B.3. Efficiency of Surprise-Guided Filter

Figure 9 illustrates the efficiency of our surprise-guided filtering approach by comparing the $O(n \log n)$ sensor-selection strategy against the exhaustive 2^N subset search across the three CARLA tasks—Right Turn, Four-Lane, and Stop Sign—under four corruption types: Gaussian, Glare, Jitter, and Lag. The specific perturbations applied to the camera view are defined as follows:

- *Gaussian Noise*: adds pixel-level random variations, modeling low-light sensor interference.
- *Glare*: causes extreme overexposure that whitens the whole frame, erasing most visual information.
- *Jitter*: applies random shifts in contrast or brightness, simulating unstable or high-noise sensor conditions.
- *Latency (Lag)*: causes the observation to lag behind the true state by reusing stale frames for several steps, emulating delayed or frozen sensor updates.

The evaluation spans a wide range of augmentation intensities to test how each method scales as visual degradation increases. For Gaussian Noise and Jitter, both methods display non-monotonic trends: mild noise can be mitigated by masking high-surprise views, while stronger perturbations introduce sharp drops, particularly in the Four-Lane task at intensity level 4. Glare yields more stable performance because overexposed channels are consistently identified and suppressed. Lag produces a near-monotonic decline across all tasks due to its strong disruption of temporal coherence. Crucially, the green $O(n \log n)$ curves typically track the red 2^N curves closely, despite the exponential cost associated with the brute-force method. This demonstrates that the surprise-guided filter attains robustness comparable to exhaustive subset evaluation at only a fraction of the computational burden. The results confirm that the learned surprise signal provides a reliable ranking over sensor utility, making combinatorial search unnecessary and enabling practical, scalable deployment in multi-sensor RL settings.

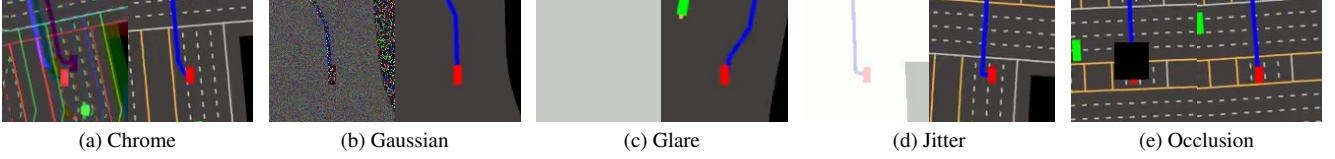


Figure 7. Visual comparison of the clean and corrupted from the CARLA BEV perspective. The left frame is the corrupted observation, and the right frame is the ground truth observation.

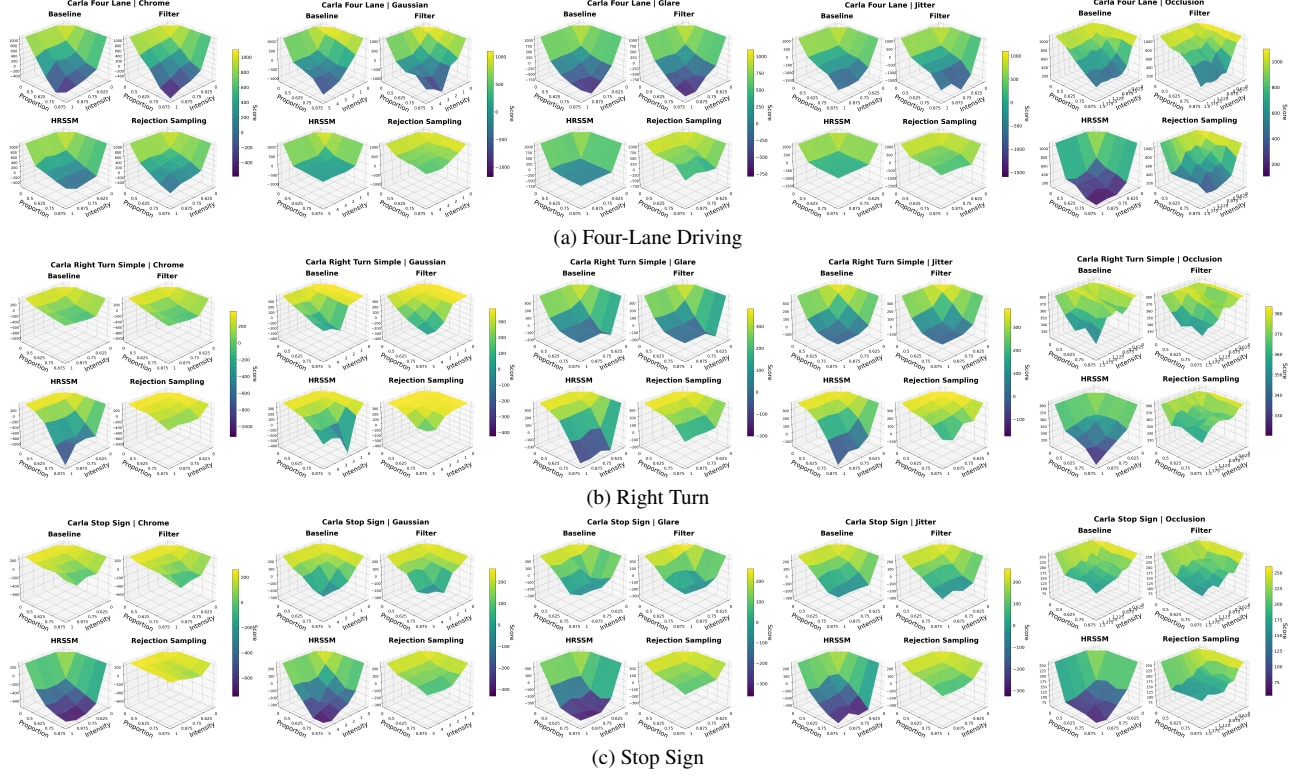


Figure 8. Qualitative results across three CARLA tasks under five corruption types: Chrome, Gaussian, Glare, Jitter, and Occlusion.

B.4. Sensor Failure Analysis

Figure 10 provides a comprehensive breakdown of how increasing sensor failures affect agent performance across all three CARLA tasks—Right Turn, Stop Sign, and Four-Lane Driving—under four primary corruption types detailed in Section B.3: Gaussian Noise, Glare, Jitter, and Latency.

Each subfigure in Figure 10 plots episodic score as a function of the number of failed sensors, comparing four methods: the Baseline world-model agent, the Augmented agent trained with Gaussian noise, the RME masked-encoder agent, and our Confident Representation approach. Across all tasks and corruption types, Baseline exhibits the steepest decline, often collapsing to near-zero or negative scores as failures accumulate. RME reduces variance but remains sensitive to structured failures such as Glare and Latency. Augmented improves stability under moderate Gaus-

sian noise, but still degrades noticeably when failures intensify.

In contrast, Confident Representation consistently maintains the highest performance curve, showing only mild degradation even with multiple failed sensors. Its advantage is particularly notable under Glare and Jitter, where corrupted visual channels severely mislead other methods. Under Latency, all agents experience some decline, yet Confident Representation preserves stable, positive performance while avoiding the abrupt failures observed in Baseline and RME.

Overall, these results show that surprise-guided representation selection remains effective across diverse CARLA configurations, enabling the agent to identify corrupted views, suppress unreliable inputs, and sustain robust control even as sensor failures accumulate.

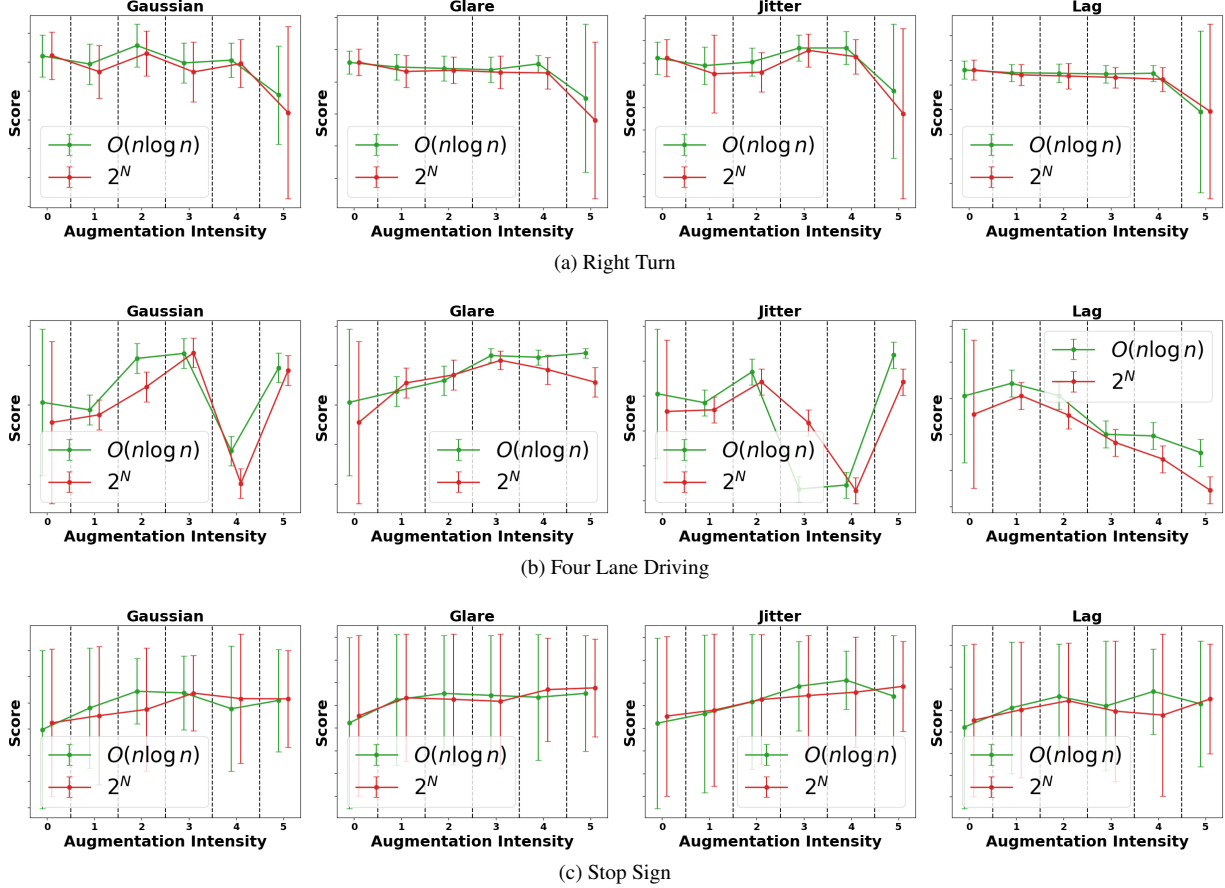


Figure 9. Comparison across three tasks under different noise conditions. We compare the 2^N brute force variant against our proposed $O(n \log n)$ Algorithm variant 2.

C. Safety Gymnasium Experiments

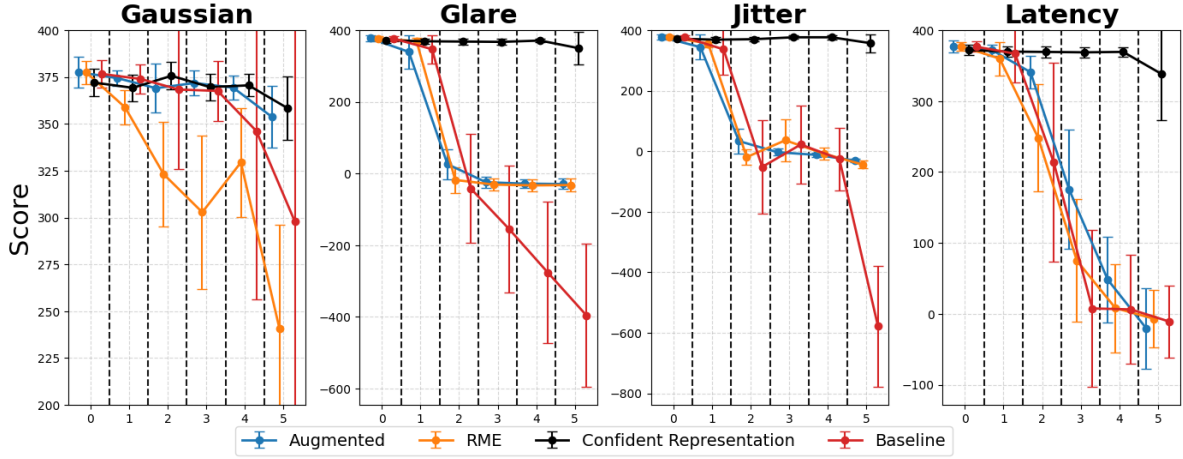
In the Safety Gymnasium domain [18], the agent’s goal is to navigate safely to objectives in an environment with obstacles. We provide three overlapping camera views as described in Section A, which capture similar scene content from complementary perspectives. We evaluate performance on three tasks: PointGoal1, where a point-mass agent must reach a goal location while avoiding hazards; PointButton1, where a point-mass must press buttons scattered in the map while avoiding obstacles, and CarGoal1, where a car-like agent must reach a goal on a road with obstacles. Each task also penalizes unsafe collisions with a cost.

Figure 3 in the main text summarizes performance under four visual corruption types—Gaussian, Jitter, Glare, and Occlusion—applied independently to each of the three camera views. Since these views are partially redundant, corruption in a single view does not immediately prevent task completion but can mislead the policy when the corrupted view dominates the latent inference. The surprise-

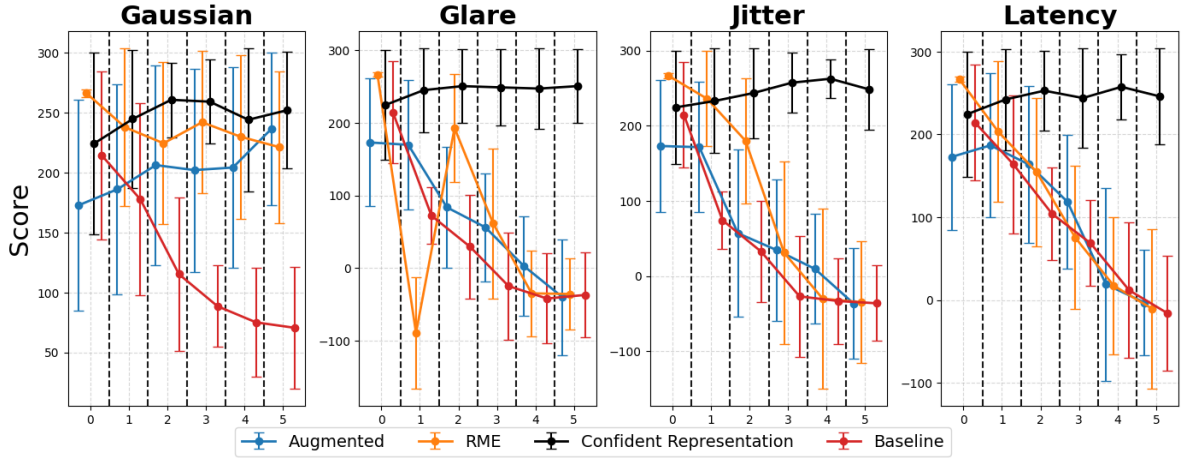
guided filter addresses this by downweighting or suppressing views whose inferred representations deviate strongly from the world model’s predictive prior. As shown in the figure, this leads to consistent improvements in both performance and the Score–Cost ratio across all Safety Gymnasium tasks. In particular, the method effectively handles Glare and Occlusion, where one view becomes unreliable but the remaining views still convey meaningful structure. The resulting gains in safety and stability demonstrate that even in environments with overlapping visual inputs, surprise-guided representation selection provides measurable robustness against view-specific corruptions.

D. Hardware Details and Hyperparameters

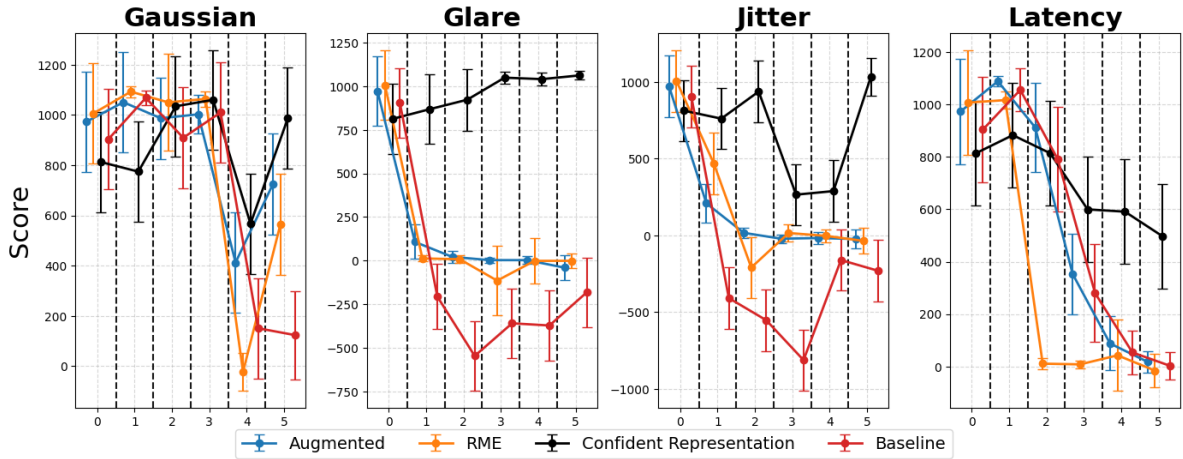
Crafter All experiments can be reproduced on a system equipped with two NVIDIA GeForce GTX 1080 GPUs with 8 GB VRAM each to handle environment complexity, an AMD Ryzen 5 5600X 6-core processor, and at least 500 MB of storage for files (excluding training data, which depends on the environment and model hyperparameters).



(a) Right Turn



(b) Stop Sign



(c) Four-Lane Driving

Figure 10. Agent performance as sensor failures increase.

Name	Symbol	Value
General		
Replay capacity	—	5×10^6
Batch size	B	16
Batch length	T	64
Activation	—	RMSNorm + SiLU
Learning rate	—	4×10^{-5}
Gradient clipping	—	AGC(0.3)
Optimizer	—	LaProp($\epsilon = 10^{-20}$)
World Model		
Reconstruction loss scale	β_{pred}	1
Dynamics loss scale	β_{dyn}	1
Representation loss scale	β_{rep}	0.1
Latent unimix	—	1%
Free nats	—	1
Actor Critic		
Imagination horizon	H	15
Discount horizon	$1/(1 - \gamma)$	333
Return lambda	λ	0.95
Critic loss scale	β_{val}	1
Critic replay loss scale	β_{repval}	0.3
Critic EMA regularizer	—	1
Critic EMA decay	—	0.98
Actor loss scale	β_{pol}	1
Actor entropy regularizer	η	3×10^{-4}
Actor unimix	—	1%
Actor RetNorm scale	S	$\text{Per}(R, 95) - \text{Per}(R, 5)$
Actor RetNorm limit	L	1
Actor RetNorm decay	—	0.99

Table 2. Hyperparameters of the Dreamerv3 model. We use the exact same parameters as discussed in [13].

CARLA and Safety Gymnasium All reported experiments related to Safety-Gymnasium and CARLA domains can be conducted on a workstation equipped with an NVIDIA GeForce RTX 4090 GPU (24 GB VRAM), an AMD Ryzen 9 7950X (16-core, 32-thread) CPU, 128 GB DDR5 RAM, and running Ubuntu 22.04 LTS (CUDA 12.4).

Cosmos All reported experiments relating to the Cosmos world model can be conducted on a workstation equipped with dual Intel Xeon 6737P CPUs (64 cores total), 2.0 TiB of DDR5 system memory, a NVIDIA RTX PRO 6000 utilizing driver version 580.82.07 (CUDA 13.0).

E. Further Results

E.1. Cosmos World Model

To further verify that the proposed rejection-based filtering mechanism generalizes beyond latent-state world models, we abstract the process of Figure 5 and extend our study to the Cosmos world model—a Diffusion-based video world model that performs pixel-space prediction through large-scale diffusion and spatio-temporal self-attention. Unlike the Dreamerv3 architecture used in our main experiments, which relies on a Variational Autoencoder with explicit la-

tent variables and KL-based surprise, Cosmos models dynamics directly in the image domain via autoregressive diffusion-based video prediction. This enables high-fidelity frame synthesis but also makes it highly sensitive to corruption in its first-frame conditioning.

We evaluate Cosmos on the robot pouring video, where a robotic manipulator pours liquid between jars on a tabletop. In our experiments, 75% of the input video is randomly perturbed with one of the noises discussed in B.2. Because Cosmos conditions its rollout on the last N frames of the input video, this corruption propagates through immediate subsequent predictions, as shown in the top row of Figure 11. After enabling our rejection mechanism, the system detects the corrupted immediate next frame as high-surprise and instead relies on internally predicted latent trajectories to reconstruct the scene. The bottom row shows markedly improved geometry and color fidelity around the robot arm and jars, indicating a safer frame to accept, demonstrating that selective rejection of corrupted conditioning frames can effectively suppress artifact propagation even in Transformer-based video world models.

E.1.1. Video Prompt: Robotic Arm Pouring Sequence

Scene Description

A robotic arm, primarily white with black joints and cables, is shown in a clean, modern indoor setting with a white tabletop. The arm, equipped with a gripper holding a small, light green pitcher, is positioned above a clear glass containing a reddish-brown liquid and a spoon. The robotic arm is in the process of pouring a transparent liquid into the glass. To the left of the pitcher, there is an open jar with a similar reddish-brown substance visible through its transparent body. In the background, a vase with white flowers and a brown couch are partially visible, adding to the contemporary ambiance. The lighting is bright, casting soft shadows on the table. The robotic arm’s movements are smooth and controlled, demonstrating precision in its task. As the video progresses, the robotic arm completes the pour, leaving the glass half-filled with the reddish-brown liquid. The jar remains untouched throughout the sequence, and the spoon inside the glass remains stationary. The other robotic arm on the right side also stays stationary throughout the video. The final frame captures the robotic arm with the pitcher finishing the pour, with the glass now filled to a higher level, while the pitcher is slightly tilted but still held securely by the gripper.

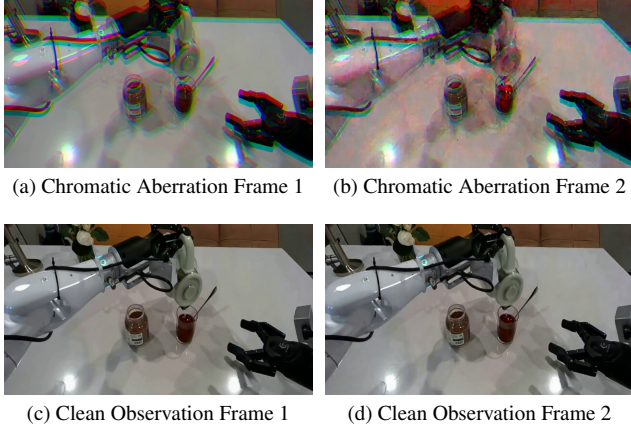


Figure 11. First-frame conditioning comparison under the Cosmos model: the first column shows the input frame, and the second column shows the corresponding generated frame that was conditioned by the first frame. We utilize the Mean Squared error between two frames as our rejection score $M(x_*)$.

E.1.2. Cosmos Evaluation Metrics

We evaluate the denoised and base Cosmos generations using metrics from the PAI-Bench [35], a comprehensive benchmark for assessing physical-world video generation and prediction quality. Following the official protocol, we use the last frame of the ground truth clean input video as the reference image to compute the image-to-video (i2v) fidelity scores. The metrics are described as follows:

- **Aesthetic Quality.** Measures the visual appeal and perceptual realism of the generated video, including composition, color balance, and absence of artefacts.
- **Background Consistency.** Evaluates the temporal stability of background regions across frames—high scores indicate minimal flicker or spatial drift.
- **Imaging Quality.** Quantifies camera-like clarity, sharpness, and signal-to-noise ratio of the generated video frames.
- **Motion Smoothness.** Assesses the temporal coherence of motion; low values indicate jitter or unnatural frame transitions.
- **Overall Consistency.** Aggregates spatial and temporal coherence into a single holistic quality measure of the entire video.
- **Subject Consistency.** Examines whether the primary object or agent remains stable in shape, color, and identity over time.
- **i2v Background.** Image-to-video fidelity for background regions, comparing each frame to the last frame of the input video.
- **i2v Subject.** Image-to-video fidelity for the subject region, capturing how faithfully the generated subject matches the reference frame.

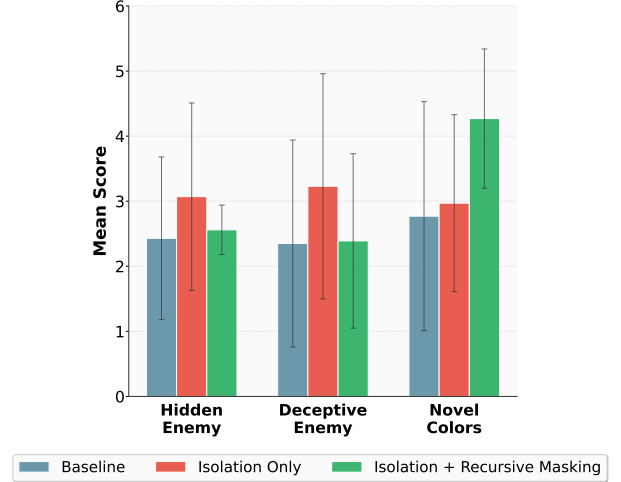


Figure 12. Crafter: Baseline vs Confident Representation Performance. Scores improve as the agent utilizes masking in Algorithm 2 to adapt to different types of novelties.

Higher scores across these dimensions indicate smoother, more coherent, and visually faithful video synthesis. In our experiments shown in Table 1, the proposed rejection sampling consistently improves both the i2v metrics and overall perceptual quality compared to the baseline generations.

E.2. Semantic Noise Experiments

Although outside the main scope of our work, we briefly consider experimentation with semantic noise.

We choose the Crafter domain [12] due to its utility in embedding abstract skins and unknown colors during inference time, primarily meant to distract the agent.

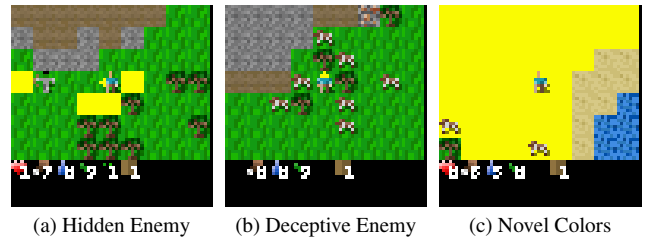


Figure 13. Qualitative visualization of agents' behavior in different semantic novelty scenarios. Figure 13a shows the *Hidden Enemy* scenario, where a yellow blob attacks the agent like zombies despite appearing harmless on the danger map. Figure 13b shows the *Deceptive Enemy* scenario, where certain cows are re-programmed to behave like zombies. Figure 13c shows the *Novel Colors*, in which all grass blocks turn bright yellow, creating a purely visual domain shift.

To furnish the agent with a rich state representation, we provide and train with six complementary observation channels (visualizations provided in Figure 14):

- *Bird’s Eye View*: A top-down, default view of the Crafter gameplay centered on the agent and a few blocks around it.
- *Grayscale*: A grayscale version of Bird’s Eye View.
- *Semantic View*: A top-down view of the entire game map showing object locations; built-in Crafter.
- *Danger Heatmap*: Shows sources of danger (zombies, skeletons, lava) as red with Gaussian decay; player represented with a green dot.
- *Health Trail*: Records agent health at its location through pixel intensity, widened to a 3x3 region to improve training.
- *Proximity Grid*: A 3x3 grid representing the eight cardinal directions around the agent; intensities reflect the presence of objects cumulatively, scaled up to 64x64 for training.

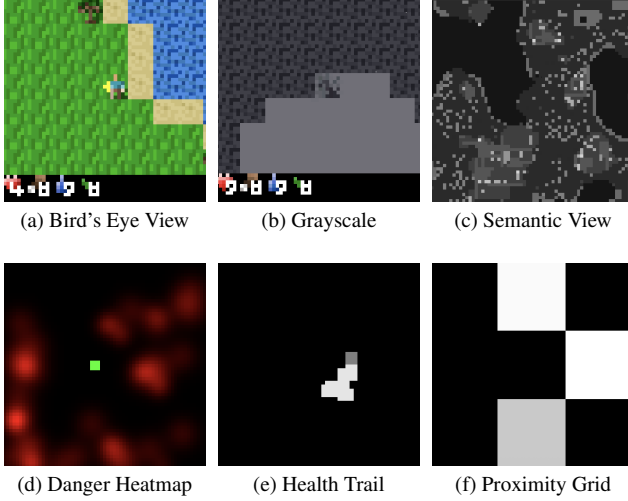


Figure 14. Six different representations of the Crafter environment, capturing a diverse range of data, were used to train the agent.

Unlike Safety Gymnasium and CARLA, the Crafter experiments use a single sensor (top down local view) but with multiple representations (grayscale, danger heatmap, proximity heatmap, etc.).

We first train the world model in the Crafter domain with Algorithm 1 to prepare for missing representations. To simulate unknown semantic noise, we inject unfamiliar contexts (invisible enemies, different enemy skins, visual color shifts) during inference time. In this setting, we create alterations that have the potential to conditionally affect multiple representations at once rather than independently (see Table 15 for a brief summary of expected representations affected given the semantic noise introduced).

We monitor the performance of a (1) Base agent, (2) an agent that only implements Step 1 *Isolation* (an agent only selects a single representation) and (3) an agent that imple-

Novelty	BEV	Gray	Sem	Dng	HT	Prox	Focus
Hidden	Large	Small		Large		Large	Low
Deceptive	Large	Large			Large	Large	Low
Colors	Large	Small				Small	High

Figure 15. Representation modalities affected by each novelty type. Large circles indicate representations directly modified by the novelty, while small circles indicate representations indirectly affected through downstream effects. The Focus column indicates whether the novelty’s impact is concentrated (High) or distributed across multiple modalities (Low).

ments Step 1 and Step 3 *Recursive Masking* (an agent that recursively masks the observations in order of surprise). We include hardware details in Appendix D.

We report our findings in Figure 13. We hypothesize that the level of depth generally affects the performance of the agent, depending on how focused the semantic noise is on a subset of representations. Recursive masking appears to improve the performance on changes that are expected to focus on a single representation, such as color based alterations that primarily affect the image representation of the state (Novel Colors). Whereas representation isolation tends to be more useful towards returning the agent towards a predictable policy when many of the representations are affected (Hidden Enemy, Deceptive Enemy, Invert Health). Although we primarily test with out-of-distribution sensor failures, we find agents equipped with Algorithm 2 appear to show some potential towards exhibiting improved stability through automated selection of representations in this setting, potentially opening avenues for future research.

E.3. Ablations

E.3.1. Representation Dropout Training

We further conduct an ablation study shown as Figure 16 to evaluate the effect of representation-dropout training on the stability and robustness of the world model. In this setting, a random subset of input modalities or encoded features is masked during each training step, encouraging the model to infer consistent latent dynamics even under partial observations.

Across all six tasks— PointGoal2, CarGoal1, PointButton1, Four-Lane Driving, Right Turn, and Stop Sign —we observe that the dropout-trained agents achieve comparable or slightly higher final scores than normal training while exhibiting smoother and more stable learning dynamics. Although the early stages of training progress more slowly due to random masking, the representation-dropout models converge to similar or better performance with noticeably re-

duced variance. This demonstrates that exposing the world model to incomplete or corrupted representations during training improves its ability to maintain coherent latent dynamics under missing-sensor or noisy conditions. The results validate that representation dropout enhances generalization and forms the foundation for the robust filtering and rejection mechanisms used in later experiments.

E.3.2. Alternative Rejection Scores $M(x_*)$

We test additional choices of $M(x_*)$ that a user may provide on the Carla Four Lane Task (Denoising is still determined by $D(x_*)$ used in the main results):

- **Surprise:** To decide whether to enter predictive mode, we experiment with utilizing the Bayesian surprise measure discussed in Eq 2.
- **Resnet-18 [14]:** To decide whether to enter predictive mode, we train a Binary Resnet-18 classifier on a synthetic dataset of observations (200,000) with and without Gaussian noise, the goal of the classifier is to reject or accept the observation given to the agent.

We report our results across two levels of noise *proportion* in Table 3.

Method	Gaussian	Chromatic	Jitter	Glare	Occlusion
Res-18, (.75)	571.9±223.6	-188.4 ± 359.9	-466.1 ± 474.7	-578.1 ± 633.9	483.5±285.9
Surp. (.75)	241.7±287.7	105.5±191.3	71.7±194.9	100.4±130.7	275.6±332.8
Res-18, (.875)	391.2±191.4	-439.1 ± 421.7	-1237.0 ± 638.3	-717.2 ± 622.4	228.6±199.5
Surp. (.875)	164.1±259.1	-18.14 ± 250.1	39.18±104.38	55.1±96.93	257.5±281.4

Table 3. Comparison of surprise and trained ResNet-18 rejectors across five noise types with different proportions (.75 and .875).

E.3.3. Alternative Denoising Methods $D(x_*)$

We test additional choices of $D(x_*)$ that a user may provide on the Carla Four Lane Task (Acceptance is still determined by $M(x_*)$ used in the main results):

- **Prior Interpolation:** We experiment with a linear interpolation between the world model’s expected observation and the current observation.
- **Restormer [33]:** We leverage an off-the-shelf denoiser with a transformer architecture.

We report our results across two levels of noise *intensity* in Table 4.

Method	Gaussian	Chromatic	Jitter	Glare	Occlusion
Interp (.625)	316.7±256.1	228.2±194.6	265.3±295.0	466.4±242.9	334.8±163.0
Restormer (.625)	262.7±221.1	180.1±177.3	196.1±170.1	308.8±211.7	305.7±185.4
Interp (.75)	258.3±288.6	205.6±273.4	183.8±401.4	448.2±274.3	255.9±189.9
Restormer (.75)	237.0±299.1	144.7±218.4	186.9±193.3	256.6±217.4	215.4±185.4

Table 4. Comparison of Interpolation and Restormer denoising methods across five noises types with different intensities (.625 and .75).

E.4. Additional Algorithms

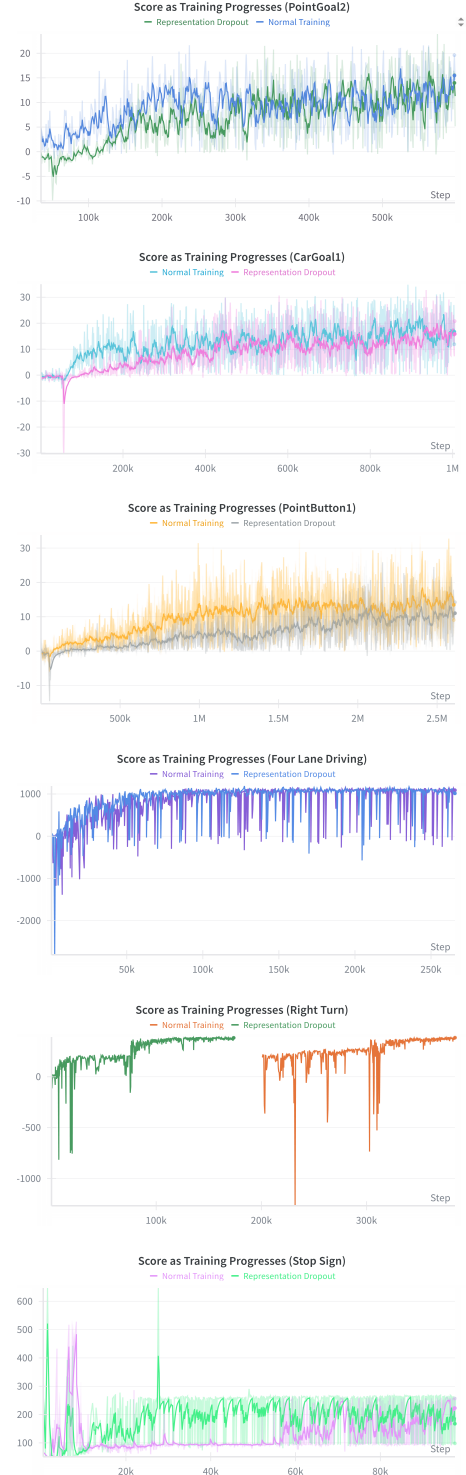


Figure 16. We compare the original world model training in Safety Gymnasium and Carla Domains to the representation dropout training.

Algorithm 1: Random Multi-Representation Dropout Training

Input: Data dictionary $\mathcal{D}$ with image keys K , mask value m

Output: Masked data dictionary $\mathcal{D}'$

Step 1: Select available images;

$K' \leftarrow \{k \in K : k \in \mathcal{D}\};$

$n \leftarrow |K'|;$

if $n = 0$ **then**

return $\mathcal{D}$

Step 2: Sample images to mask each step;

For each batch b and timestep t ;

 Draw $u_{b,t} \sim \text{Uniform}\{0, \dots, n-1\};$

Step 3: Assign random rankings to images;

For each (b, t) ;

 Draw random values $r_{b,t,1}, \dots, r_{b,t,n};$

 Compute rankings $\pi_{b,t}$ by sorting these values;

Step 4: Construct masking matrix;

For each (b, t, i) ;

 Set $\text{mask}[b, t, i] \leftarrow 1$ if $\pi_{b,t}(i) < u_{b,t}$, else 0;

Step 5: Apply masking;

For each $k_i \in K'$;

$\mathcal{D}'[k_i][b, t] \leftarrow \begin{cases} m, & \text{if } \text{mask}[b, t, i] = 1; \\ \mathcal{D}[k_i][b, t], & \text{otherwise} \end{cases};$

return $\mathcal{D}'$;

Algorithm 2: $O(n \log n)$ Surprise-Guided Representation Selection

Input: Observation dictionary obs_t , sensor keys K , world model WM, prev state z_{t-1} , prev action a_{t-1} , Optional Depth D

Output: Surprise values $\mathcal{S}$, latents $\mathcal{Z}$

Step 1: Compute surprise for each sensor individually;

foreach $k \in K$ **do**

 Initialize empty observation;

$\text{obs}_{t'} \leftarrow \mathbf{0};$

 Isolate observation sensor;

$\text{obs}_{t'}^k \leftarrow \text{obs}_{t'}^k;$

 Encode: $e_t \leftarrow \text{Encoder}(\text{obs}_{t'});$

 Predict posterior distribution;

$P_\phi(z_t^k | e_t, h_t) \leftarrow \text{WM}(z_{t-1}, a_{t-1}, e_t);$

 Surprise;

$S_k \leftarrow \text{KL} [P_\phi(z_t^k | e_t, h_t) \parallel P_\phi(z_t^k | h_t)];$

 Append S_k to $\mathcal{S}$ and $z_t^k \sim P_\phi(z_t^k | e_t, h_t)$ to $\mathcal{Z}$;

Step 2: Sort sensors by decreasing surprise;

$\pi \leftarrow \text{argsort}([S_k]_{k \in K});$

Step 3: Iteratively mask sensors in sorted order;

$\text{obs}_{t'} \leftarrow \text{obs}_t;$

for $i \leftarrow 1$ **to** $\min(|K|, D)$ **do**

 Mask sensor π_i ;

$\text{obs}_{t'}^{\pi_i} \leftarrow \mathbf{0};$

 Encode: $e_t \leftarrow \text{Encoder}(\text{obs}_{t'});$

 Predict posterior distribution;

$P_\phi(z_t^k | e_t, h_t) \leftarrow \text{WM}(z_{t-1}, a_{t-1}, e_t);$

 Surprise;

$S_k \leftarrow \text{KL} [P_\phi(z_t^k | e_t, h_t) \parallel P_\phi(z_t^k | h_t)];$

 Append S_k to $\mathcal{S}$ and $z_t^k \sim P_\phi(z_t^k | e_t, h_t)$ to $\mathcal{Z}$;

$\mathcal{Z}^* \leftarrow \arg \min_{\mathcal{Z}} \mathcal{S}(\mathcal{Z})$

return $\mathcal{Z}^*$
