2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Foundational World Models Accurately Detect Bimanual Manipulator Failures

Isaac R. Ward[†,*,1], Michelle Ho[*,1], Houjun Liu[1], Aaron Feldman[1], Joseph Vincent[1], Liam Kruse[1], Sean Cheong[2], Duncan Eddy[1], Mykel J. Kochenderfer[1] and Mac Schwager[1]

***Abstract*— Deploying visuomotor robots at scale is challenging due to the potential for anomalous failures to degrade performance, cause damage, or endanger human life. Bimanual manipulators are no exception; these robots have vast state spaces comprised of high-dimensional images and proprioceptive signals. Explicitly defining failure modes within such state spaces is infeasible. In this work, we overcome these challenges by training a probabilistic, history informed, world model within the compressed latent space of a pretrained vision foundation model (NVIDIA's Cosmos Tokenizer). The model outputs uncertainty estimates alongside its predictions that serve as non-conformity scores within a conformal prediction framework. We use these scores to develop a runtime monitor, correlating periods of high uncertainty with anomalous failures. To test these methods, we use the simulated Push-T environment and the Bimanual Cable Manipulation dataset, the latter of which we introduce in this work. This new dataset features trajectories with multiple synchronized camera views, proprioceptive signals, and annotated failures from a challenging data center maintenance task. We benchmark our methods against baselines from the anomaly detection and out-of-distribution detection literature, and show that our approach considerably outperforms statistical techniques. Furthermore, we show that our approach requires approximately one twentieth of the trainable parameters as the next-best learning-based approach, yet outperforms it by 3.8% in terms of failure detection rate, paving the way toward safely deploying manipulator robots in real-world environments where reliability is non-negotiable.**

I. Introduction

Organizations are increasingly deploying robotic systems in high-stakes environments where failures can cause unwanted delays, damage property, and risk human safety. Among these systems, bimanual manipulators—robots with two coordinated arms—are of special interest because they are designed to perform complex, human-like manipulation tasks such as assembly, tool use, or handling deformable objects. These tasks demand tight coordination between both arms, making the systems especially vulnerable to small perception or control errors. Failures in bimanual manipulation can cascade, compounding damages across fleets and ultimately impeding deployments. To achieve safe deployments, we clearly need scalable methods to reliably detect and mitigate such failures as they arise.

[†]Corresponding author.
[*]Authors contributed equally to this work.
[1]Stanford University, Stanford, California, 94305, USA. {irward, mtho, houjun, lkruse, aofeldma, lkruse, deddy, schwager, mykel}@stanford.edu
[2]Watney Robotics, San Francisco, California, 94110, USA. sean@watneyrobotics.com
*Watney Robotics provided funds to support this work.
*Toyota Research Institute provided funds to support this work.

Failures are typically associated with anomalies, but these can be challenging to define, and often, we are simply interested in flagging any behavior that is unlike a set of observed "good" or "nominal" behavior. For robots, behavior that meaningfully differs from nominal operation might include unusual sequences of visual or proprioceptive states. However, due to high data-rates and large volumes of data (e.g. modern robots use multiple 4K camera feeds at 60Hz for perception), parsing the data in real-time to detect failures is challenging. How can we efficiently represent the high dimensional behavioral sequences of visuomotor robots?

The approach taken in this work is to leverage pretrained foundation vision models to enable the training of world models in a compressed latent space. These world models learn what constitutes good behavior. World models—learned models that forecast future images or other sensory modes conditioned on an action sequence—have emerged as a powerful paradigm in robotics and physical AI. A world model can provide a robot with a means of determining the consequences of potential actions before executing them [1], [2], which the robot can use for downstream error detection and recovery, data generation for policy training, and post-failure analysis through counterfactual scenario reasoning.

We propose a probabilistic variational auto-encoder (VAE) style world model, which provides a quantifiable measure of uncertainty associated with its predictions. We learn only the nominal dynamics—that is, we train the world model entirely on nominal examples—minimizing the uncertainty of predictions while the robot is executing desired behavior. We then use the world model as a runtime monitor, proposing two different error metrics: (i) the intrinsic variance estimated by the VAE at runtime, and (ii) the empirical error obtain by comparing the forecast with the ground truth. We calibrate failure thresholds for both metrics using conformal prediction [3]. This enables us to separate anomalous failures from nominal operation at runtime.

Our contributions are thus threefold:

1) We propose a probabilistic world model trained within the latent space of NVIDIA's pretrained *Cosmos Tokenizer*[1] [2]. By leveraging this pretrained tokenizer, we create a latent space VAE-based world model with less than 600k trainable parameters.
2) We propose two methods for failure prediction with our world model: (i) a VAE uncertainty estimate, and

[1]NVIDIA's Cosmos Tokenizer is a vision autoencoder specialized for manipulator images from the Cosmos video foundation model platform.

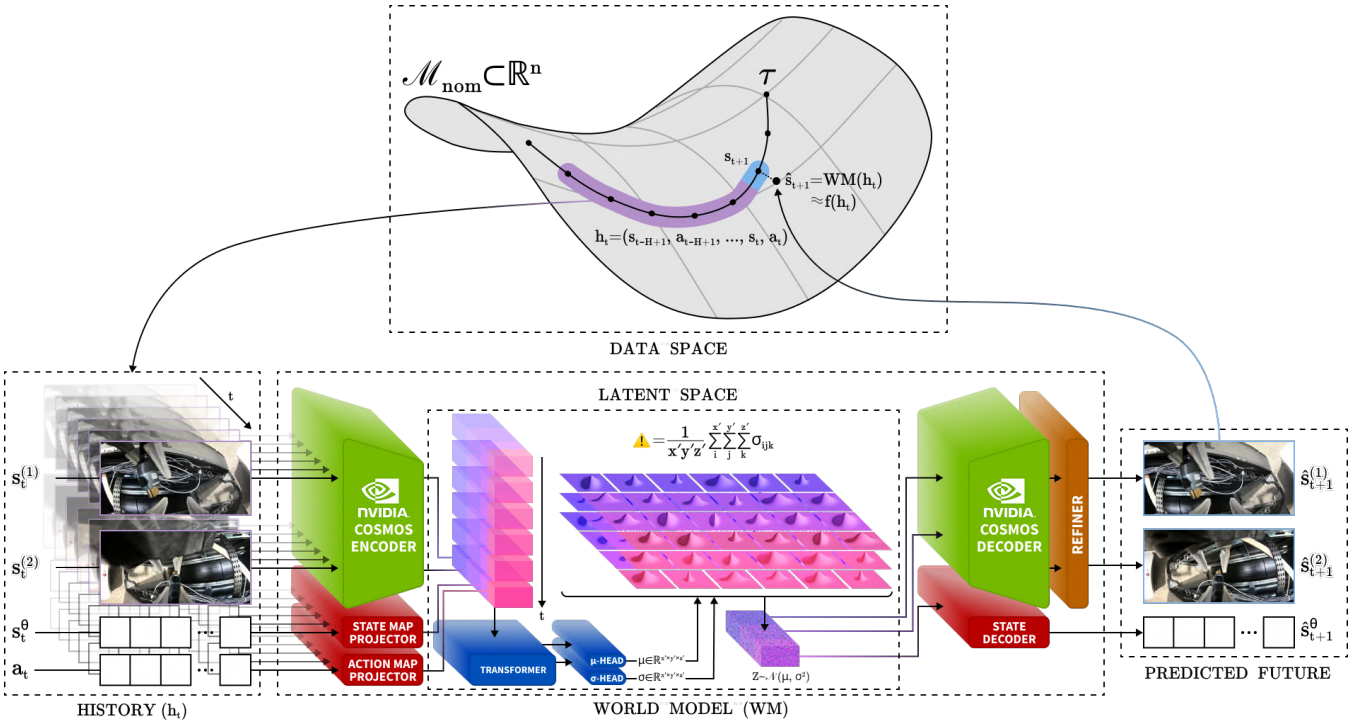


Fig. 1: A schematic of the training process for the proposed method. A trajectory $\tau = (s_1, a_1, s_2, a_2, \dots)$ is sampled from the dataset and sliced into training samples consisting of a fixed history window $\mathbf{h}_t = (s_{t-H+1}, a_{t-H+1}, \dots, s_t, a_t)$ and future state s_{t+1}. The multi-step, multi-view images captured at the robot's grippers ($s_t^{(1)}$ and $s_t^{(2)}$) are encoded using a foundation video encoder (Cosmos) and—alongside learned projections of the historical proprioceptive states s_t^θ and actions a_t—are combined into latent feature maps. These are processed sequentially by a transformer to predict a tensor of distributions, the standard deviations of which quantify the uncertainty of the future prediction. This uncertainty is lower for nominal inputs (like those observed during training), and higher for anomalous inputs associated with failures. In other words: $\texttt{WM}(h_t) \approx f(h_t)$ only when $h_t \in \mathcal{M}_{\text{nom}}$. Sampling from this latent distribution gives a latent feature map $Z \sim \mathcal{N}(\mu, \sigma^2)$, which can be decoded and refined into predictions for the next immediate state ($\hat{s}_{t+1}^{(1)}$, $\hat{s}_{t+1}^{(2)}$, $\hat{s}_t^\theta$).

(ii) an empirical forecast error, both of which outperform five baseline failure detection methods from the literature.

3) We introduce the *Bimanual Cable Manipulation dataset*, a new dataset featuring labeled nominal and failure trajectories from real world fleets of bimanual robot manipulators in a data center bring-up and maintenance task.

II. Related Work

A. Classical and Statistical Methods

Classical methods, including control charts, hypothesis testing, residual analysis, and change-point detection, are widely used across disciplines such as quality control and computer networks [4] to detect anomalies that may result in failure. They are simple in formulation and benefit from interpretability and minimal computational overhead [5], but lack the representational power and adaptability of modern learning-based techniques. Moreover, these methods are ill-suited for robotics, where strong assumptions such as stationarity, noise independence, and accurate dynamics models often break down [6]. Robotic systems are high-dimensional, multimodal, and context-dependent, with temporal correlations that classical techniques fail to capture.

Ensemble-based statistical methods attempt to address some of these shortcomings by aggregating multiple detectors, each sensitive to different statistical properties of the data. Common ensemble approaches for OOD detection include ensembles of classifiers [7] or generative models [8], and hybrid methods that incorporate Bayesian uncertainty estimation [9], [10]. These methods estimate uncertainty from limited knowledge of the data distribution, flagging high variance or predictive disagreement as anomalies.

B. Autoencoder and Embedding-Based Methods

Representation learning methods aim to discover feature spaces in which nominal and anomalous states can be more easily separated. Autoencoder-based approaches learn these feature spaces via self-supervised reconstruction, with reconstruction error serving as an anomaly score [11], [12].

Similarly, embedding-based methods operate on the assumption that in-distribution data lies in compact regions of feature space, while anomalies lie farther from the nominal manifold [5]. Non-conformity scores are then defined in terms of distance or similarity between a test embedding

and nominal embeddings. Parametric approaches include Mahalanobis distance, which assumes Gaussian structure in the embedding space [13], while non-parametric approaches rely on neighborhood similarity [14], [15]. Embedding-based approaches are model-agnostic, and integrate naturally with supervised or self-supervised representation learning. Despite their flexibility, embedding-based methods depend heavily on the quality of the learned feature space, limiting reliability under distribution shifts.

C. Distribution Models

Distribution modeling approaches explicitly attempt to learn the nominal data distribution, with the assumption that out-of-distribution (OOD) samples will have low likelihood under the model. Variational autoencoders [16], autoregressive models [17], and normalizing flows [18], [19] have all been applied to this problem. However, it has been shown that such models can assign high nominal likelihoods to anomalous data [20], particularly in the case of normalizing flows, which tend to capture low-level pixel correlations rather than high-level semantic features [21]. These methods are also computationally demanding, requiring large amounts of diverse nominal data for training and significant resources for inference.

D. Modern methods

Some recent works have proposed to use world models, such as DINO [22] for error recovery [23] and human intent alignment [24] in visuomotor robot policies. Generally speaking, world models learn from a training dataset how a robot's state evolves given a sequence of actions. When test-time inputs deviate from the training dataset, changes in the world model's confidence or epistemic uncertainty allow for OOD detection [25], [26].

Finally, large language model (LLM)-based approaches transform visual or sensory inputs into symbolic or natural language descriptions, which are then evaluated for plausibility by an LLM [27], [28]. At a semantic level, this mirrors human anomaly detection, where context and meaning guide judgments rather than low-level patterns. Such approaches promise better generalization to novel scenarios, interpretability through natural language explanations, and the integration of common-sense reasoning. However, challenges remain, including sensitivity to prompt design, latency, and susceptibility to hallucination or bias. While still nascent, these methods highlight the potential of foundation models to reason about anomalous behavior in robotics.

III. Problem Formulation

We consider the problem of identifying trajectories τ that are considered anomalous. We assume nominal behavior trajectories approximately lie on a lower dimensional manifold $\mathcal{M}_{\text{nom}} \subset \mathbb{R}^n$ within the n-dimensional data space such that $\dim(\mathcal{M}_{\text{nom}}) = m < n$ (see Figure 1). Any $\tau \in \mathcal{M}_{\text{nom}}$ that remains close to $\mathcal{M}_{\text{nom}}$ is considered nominal, while trajectories that deviate significantly are treated as anomalous failure trajectories.

The system is governed by unknown dynamics f

$$s_{t+1} = f(\mathbf{h}_t), \tag{1}$$

where $s_{t+1} \in \mathbb{R}^n$ is the next state and $\mathbf{h}_t$ is the fixed-length history window of H past state-action pairs at time t

$$\mathbf{h}_t = (s_{t-H+1}, a_{t-H+1}, \ldots, s_t, a_t) \in \mathbb{R}^{H\times(n+d)}. \tag{2}$$

Our objective is then to construct a classifier function

$$C : \mathbb{R}^{H\times(n+d)} \to \{0, 1\}, \tag{3}$$

such that

$$C(\mathbf{h}_t) = \begin{cases} 0, & \text{if } \mathbf{h}_t \text{ is a trajectory near } \mathcal{M}_{\text{nom}}, \\ 1, & \text{otherwise,} \end{cases} \tag{4}$$

where $C = 0$ indicates nominal behavior and $C = 1$ indicates anomalous behavior.

IV. Methodology

Our methodology centers on training a probabilistic, history-conditioned world model (`WM`) in the latent space of NVIDIA's Cosmos Tokenizer and then using its uncertainty estimates as non-conformity scores within a conformal prediction framework. In addition, we benchmark a set of alternative non-conformity scores adapted from the anomaly detection and OOD detection literature.

A. Training the World Model

Following the schematic in Figure 1, the `WM`'s input is a history of visual observations, proprioceptive states, and actions. Raw images are first embedded using NVIDIA's Cosmos Tokenizer, yielding latent feature maps that are fused with proprioceptive and action embeddings. A transformer-based sequence model is then trained to predict distributions over future latent feature maps.

The `WM` is trained on ($\mathbf{h}_t$, s_{t+1}) pairs sampled from only nominal trajectories until the validation loss stops improving. During training, it learns to predict $\texttt{WM}(\mathbf{h}_t) = s_{t+1}$ using a reconstruction loss term that 1) maximizes perceptually accurate reconstructions in pixel space across all camera views (using a perceptual loss L_{V} [29]) and 2) minimizes the mean square error between the predicted proprioceptive state $\hat{s}_t^\theta$ and the ground truth proprioceptive state s_t^θ)

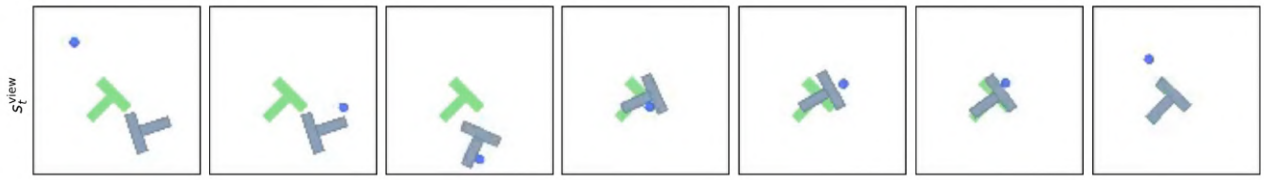
$$\mathcal{L}_{\text{recon}} = \Big[\frac{1}{N_{\text{views}}} \sum_{i=1}^{N_{\text{views}}} \mathcal{L}_{\text{V}}\big(s_{t+1}^{(i)},\, \hat{s}_{t+1}^{(i)}\big)\Big] + \frac{1}{2}\, \mathcal{L}_{\text{MSE}}\big(s_t^\theta,\, \hat{s}_t^\theta\big) \tag{5}$$

a term that minimizes reconstruction error in the latent space

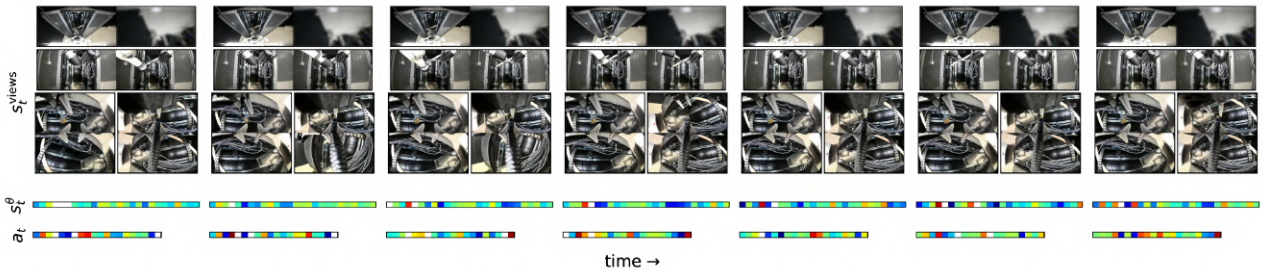
$$\mathcal{L}_{\text{recon}}^z = \mathcal{L}_{\text{MSE}}\big(z,\, \hat{z}\big) \tag{6}$$

a Kullback-Leibler divergence term that ensures that the latent distribution over future states maintains a zero mean and unit standard deviation

$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}\Big(\mathcal{N}(\mu, \sigma^2) \,||\, \mathcal{N}(0, 1)\Big) \tag{7}$$



(a) Each trajectory in the Push-T dataset consists of a 25fps, $128 \times 128 \times 3$ single-view RGB video showing a blue dot (the agent) attempting to push a T-shaped object into a green T-shaped slot. The state s_t^θ is the (x, y) position of the end effector, and the action a_t (not visualized) specifies the next change in position of the end effector.



(b) An illustration of a sequence from a single trajectory from the Bimanual Cable Manipulation dataset. The state comprises 8 camera views—each $\sim$ 60fps, $1280 \times 720 \times 3$ videos captured from the head (front and back), chest (left and right), left gripper (below and above), and right gripper (below and above) cameras—as well as proprioceptive data at each timestep $s_t^\theta \in \mathbb{R}^{52}$. Action vectors $a_t \in \mathbb{R}^{41}$ contain motor torque and gripper activation commands at each timestep. In this work, only the gripper cameras are used as the cable is most visible from these views.

Fig. 2: Overview of datasets used in this work: (a) Push-T dataset, (b) Bimanual Cable Manipulation dataset.

and a negative log likelihood term that ensures that the ground truths would be likely under the learned latent distribution

$$\mathcal{L}_{\text{NLL}} = -\,\mathbb{E}_{z\sim\mathcal{N}(\mu,\sigma^2)}\big[\log\mathcal{N}(z;\mu,\sigma^2)\big] \tag{8}$$

and they are weighted as follows to form the total loss used during training and validation

$$\mathcal{L} = {}^1\!/_{10}\,\mathcal{L}_{\text{recon}} \;+\; 2\,\mathcal{L}^z_{\text{recon}} \;+\; {}^1\!/_{20}\,\mathcal{L}_{\text{KL}} \;+\; \mathcal{L}_{\text{NLL}} \tag{9}$$

Training begins with a one-step prediction objective, and every 16 epochs, the autoregressive prediction horizon is doubled as part of a curriculum learning strategy, until a maximum horizon of 32 steps is reached. This gradual extension of the rollout length stabilizes training and encourages the model to capture longer-term dynamics.

B. Non-conformity scores

To detect anomalous failures, we need to translate the world model prediction into a non-conformity score. A non-conformity score is a scalar that quantifies how 'unlike' the nominal training set the robot's behavior is at a given timestep. Different scoring methods define this non-conformity with respect to different properties. We investigate the following non-conformity scores for classifying nominal from failure behavior.

1) `WM` uncertainty: defined as the average of each of the standard deviations in the normal distributions over future latents (see Figure 1). For a 3D tensor of distributions with dimensions $x' \times y' \times z'$ the `WM` uncertainty score is $\frac{1}{x'y'z'}\Sigma_i^{x'}\Sigma_j^{y'}\Sigma_k^{z'}\sigma_{ijk}$.

2) `WM` prediction error: defined as the discrepancy between the world model's predicted next-step states and the actual observed states. We compute this in latent space (using Equation 6), reflecting mismatches in compressed features.

3) `logpZO` [19]: refers to the log-probability $-\log p_{\text{flow}}(z_{t+1})$ under a Zero-order (ZO) normalizing flow model trained on nominal latent trajectories. Low log-probability values indicate that an input is unlikely under the nominal distribution.

4) `AE` reconstruction error [21], [30]: the anomaly score is the mean squared error $\|s_t - \hat{s}_t\|_2^2$ between the input s_t and reconstruction $\hat{s}_t$ using an autoencoder trained entirely on nominal states. Failures tend to yield higher reconstruction errors than nominal inputs.

5) `AE` sim(z, z_{safe}) [28]: the similarity in latent space between a test embedding z and the nearest embedding from a known safe set of nominal data z_{safe}. Inputs far from the nominal safe set are flagged as anomalous. We use mean-squared error as our similarity function.

6) `SPARC` [31]: measures the smoothness of a trajectory by looking at the arc length of the Fourier magnitude spectrum of sequential states. A smoother trajectory has less high-frequency content (shorter spectral arc length), while a jerkier trajectory has more high-frequency oscillations (longer arc length). This scalar smoothness measure is used as the non-conformity score.

7) `PCA K-means` [32]: a PCA-based dimensionality reduction followed by K-means clustering into two groups (nominal and failure). The non-conformity score is the distance from the nearest cluster centroid $\min_{c\in\{1,2\}} \|z - \mu_c\|_2$.

This is the only method that we benchmark that requires failure data.

8) Random: a random baseline that uniformly assigns non-conformity scores at each timestep in the range $[0, 1]$ at random, regardless of the information available. This provides a lower bound equivalent to chance-level classification.

C. Conformal prediction

We use conformal prediction (CP) to calibrate thresholds for each non-conformity score [33]. For every trajectory, the score sequence is convolved with a uniform triangular filter that most heavily weights the most recent samples, with window length 50, which has the effect of smoothing out high frequency aberrations. The maximum value is taken as the trajectory-level statistic.

Thresholds are fit to these trajectory-level statistics using only a held-out set of nominal trajectories, with no access to failure data. At test time, a trajectory is flagged anomalous if its statistic exceeds the $(1 - \alpha)$ quantile of the nominal calibration distribution, giving a guaranteed false alarm rate of at most α. To reduce bias from how the calibration set is selected, we use a delete-d jackknife with 32 random permutations, where in each permutation we fit thresholds on half of the held-out nominal trajectories and evaluate on the remaining half, before averaging the resulting thresholds. Table I and Figure 5 show these results.

There are some caveats to using this approach. The validity of these guarantees relies on the assumption that calibration and test data are exchangeable, i.e., the probability of a sequence of values is independent of the order of the sequence. The time series data produced in real-world robotic deployments tends to be temporally correlated and therefore not exchangeable. However, we mitigate this by using a summary statistic per trajectory, meaning that this condition is met if trajectories are exchangeable, even if timesteps within trajectories are not. Furthermore, changing environment conditions, operator behavior, hardware wear, or sensor drift tend to shift the data distribution, so conformal guarantees on the calibration data do not strictly hold for deployment data. In such cases the nominal thresholds may misestimate error rates, highlighting the need for adaptive calibration or re-fitting in long-term deployments.

V. Results & Discussion

A. The Push-T dataset

We use the Push-T environment to create a dataset to test our WM method. The Push-T environment tasks an end-effector agent controlled by a visuomotor diffusion policy [34] to push a blue T-shaped object into a green T-shaped space (Figure 2). We generate 1028 nominal rollouts for training, and 128 nominal rollouts for validation (each with different initialization seeds). We generate a further 512 nominal and 2048 failure rollouts (512 each across 4 distinct failure modes) for evaluating our methods. The four modes each have a unique disturbance which we apply to force failure: changing the T from blue to orange, changing the T from blue to green, reducing the friction of the T, and removing friction altogether.

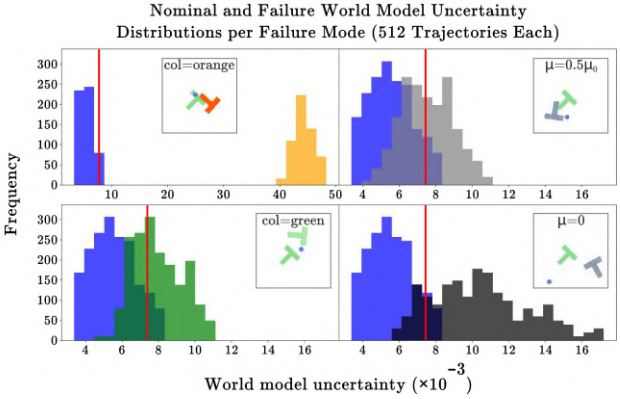


Fig. 3: Histograms of the WM uncertainty scores for nominal (blue), anomalous recolored, and anomalous dynamically altered inputs, with overlaid visualizations of what these anomalous inputs look like. In each example the 90% conformal prediction threshold is rendered in red. To test our method's ability to handle visual and proprioceptive anomalies, we alter the color to orange (top left) and green (bottom left) as well as altering the friction to half the nominal friction coefficient (top right) and zero (bottom right).

As outlined in Figure 3, we find that the WM uncertainty scores for the OOD inputs are generally higher than the scores for the nominal trajectories across multiple OOD modes. In particular, we find that our method—which ingests both visual and kinematic data—is able to detect visual anomalies (changing colors) and changes to the dynamics (altered friction), clearly separating nominal and anomalous behavior.

B. The Bimanual Cable Manipulation dataset

We introduce the Bimanual Cable Manipulation dataset to further test our approach. The dataset features 83 nominal trajectories for training and validation, 7 held-out nominal trajectories for fitting conformal prediction thresholds, 7 held-out nominal trajectories for evaluation, and 9 held-out failure trajectories for evaluation. The primary failure mode in this dataset is the cable being dropped by the robot's manipulators as it attempts to plug in the cable (as illustrated in Figure 7).

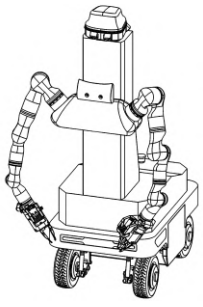


Fig. 4: The WR1 Data Center Mobile Manipulator robot.

The dataset was collected by a WR1 robot (see Figure 4) in a data center environment operating with a manual policy steered by a trained data center technician via a secured remote link from over 7000 miles away. The robots operate using a segregated private network link and proprietary secured communication protocol enabling ultra low latency communication for remote inference and operation.

Each trajectory is between 18 seconds and 3 minutes long, and features four synchronized $\sim$ 60fps, $1280 \times 720 \times 3$ RGB video feeds from cameras mounted on the robot at head height, chest height, and on the end of each manipulator (as illustrated in Figure 2). Moreover, each state also features a 52-dimensional proprioceptive state vector and 41-dimensional action vector.

In Table I (and Figure 5), we benchmark each of the methods described in Section IV-B on the Bimanual Cable Manipulation dataset using the method described in Section IV-C to fit classification thresholds. We find that our methods consistently outperform competing learning-based methods despite having only $\sim 1/20^{th}$ of learnable parameters (569.7k compared to approximately 10M) due to being trained in the latent space of a video foundation model. Moreover, our methods categorically outperform statistical methods in terms of classification accuracy.

TABLE I: Classification performance of different methods on the Bimanual Cable Manipulation dataset, using an 85% conformal prediction threshold.

	Avg. Classification Accuracy (32 folds, %)		
Method	Nominal ($N_{\checkmark}$=7)	Failure ($N_{\times}$=9)	Weighted total (N=16)
WM uncertainty **(ours)**	87.9 ± 17.0	$\mathbf{95.1 \pm 5.5}$	$\mathbf{92.0 \pm 6.4}$
WM pred. error **(ours)**	$\mathbf{88.3 \pm 17.8}$	87.5 ± 12.3	87.9 ± 6.4
logpZO	86.8 ± 20.4	91.3 ± 6.7	89.3 ± 6.8
AE recon. error	80.6 ± 20.5	45.8 ± 18.6	61.0 ± 4.2
AE sim(z, z_{safe})	80.7 ± 20.3	55.2 ± 21.4	66.4 ± 6.1
SPARC	64.7 ± 35.6	25.3 ± 18.1	42.6 ± 6.8
PCA K-means	66.9 ± 33.7	34.4 ± 8.9	48.6 ± 12.6
Random	55.3 ± 34.9	25.7 ± 20.5	38.7 ± 6.4

Interestingly, we find that the empirical coverage of our conformal prediction bounds in the evaluation set does not always match the presumed coverage set by the $(1-\alpha)$ quantile (85% in practice), implying that the evaluation data and fitting data do not necessarily come from the same underlying data distribution (see Section IV-C for further discussion of the assumptions relating to conformal prediction). We expect that this effect is due to the relatively small size of the total nominal set; it is possible to repeatedly select a biased fitting subset of the total nominal set that is meaningfully different from the remaining evaluation subset, making it possible to not observe an $\sim 85\%$ nominal classification accuracy.

Note that Random in this case does not mean randomly guessing the class, but randomly assigning a non-conformity score to each timestep of the trajectory, and then applying conformal prediction to a class-imbalanced evaluation dataset, hence the $< 50\%$ class-weighted accuracy on a binary classification task.

C. The WM uncertainty score is correlated with approaching or ongoing anomalous failures

Figure 7 outlines an unseen failure case where the WM uncertainty metric clearly increases and decreases as inputs become more or less correlated with moments of impending or present failures. Moreover, although we only classify two regimes; nominal and failure, we do observe distinct regions with different amounts of relative 'safety' for the robot. In particular, we observe that periods where the robot is not holding a cable—and it is thus impossible to drop the cable (failure)—are associated with the lowest WM uncertainty. Moreover, the score rapidly increases prior to the cable being dropped, despite the cable still visibly being grasped by the grippers. We hypothesize that this is due to the world model correlating certain proprioceptive state/action sequences with nominal behavior, and not observing such inputs in these pre-failure timesteps.

D. WM uncertainty is a better predictor of anomaly than WM prediction error

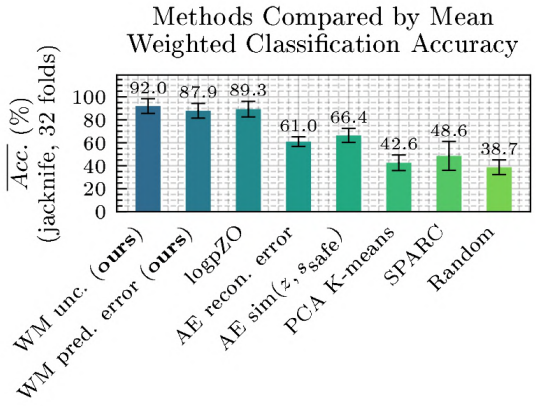


Fig. 5: We find that our WM-based approaches outperform competing methods on the Bimanual Cable Manipulation dataset, presumably because of how they uniquely incorporate historical context. See Table I for a class-based breakdown.

Figure 5 illustrates that our WM methods outperform competing methods on the Bimanual Cable Manipulation dataset, though we note that WM uncertainty is a more reliable non-conformity score than WM prediction error. We expect that this is due to the prediction error on a single outcome being less informative than the model's own distributional uncertainty, since a low-error match can occur by chance even when the input is off-manifold, whereas high predictive variance more reliably implies anomalous inputs.

E. Learning-based methods are slower but can still be executed in real-time

We benchmark the runtime of each method by measuring the wall-clock time to compute a single scalar non-conformity score. As illustrated in Figure 6, all methods

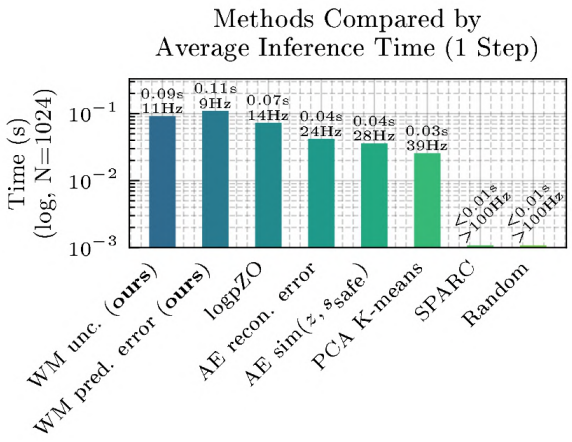


Fig. 6: The time taken to generate a scalar non-conformity score value for a single step, for each method.

operate comfortably above 9Hz, satisfying the threshold for real-time execution in our robotic setting. Statistical baselines such as `SPARC` and `PCA K-means` are the fastest, while deep learning-based methods are slower due to the forward passes through neural architectures. Among the learning-based approaches, scores that operate directly in latent space (`AE` sim(z, z_{safe})) are faster than their counterparts requiring reconstruction to pixel space (`AE` reconstruction error), as decoding latent codes requires further processing. Although `WM`-based methods are the slowest overall, their throughput remains sufficient for deployment, and world models do provide ancillary utility in a robotics context (e.g. enabling planning).

VI. Conclusion

We present a general method for detecting anomalous failures using a probabilistic world model trained entirely on data featuring nominal task completions. We find that world model uncertainty is a reliable indicator of anomalous behavior, effectively distinguishing nominal from failure trajectories in real-time, in a real-world, bimanual robot manipulator setting.

Limitations: Our approach uses conformal prediction, but violates formal assumptions on exchangeability. Nonetheless, our results show empirically that the conformal calibration is effective. Our method may flag false errors due to benign distribution shifts (e.g., background color changes). The method also inherits currently unknown biases from the pretrained tokenizer. We expect that these limitations could be resolved with more data, and by testing the efficacy of other pretrained encoders as they become available.

Future Work: Promising avenues for future work include exploring longer and more compact historical representations to improve detection of longer-term task-progression failures, conducting feature importance analysis to quantify the contribution of proprioceptive and visual data to the `WM` uncertainty metric, and finally, extending the approach to fully autonomous manipulation policies. In particular, leveraging the `WM` for simultaneous failure detection *and correction* using optimization procedures that solve for the action sequence that minimizes `WM` uncertainty may provide a solution that neatly improves the robustness of bimanual manipulators.

Acknowledgments

The authors thank Watney Robotics for providing funds, access to robot hardware, and datasets which assisted the authors with their research.

The authors thank Toyota Research Institute (TRI) for providing funds to assist the authors with their research, but this article solely reflects the opinions and conclusions of its authors and not TRI or any other Toyota entity.

Generative AI tools and technologies were used in this work to identify related literature (ChatGPT, Perplexity AI), sense-check concepts (ChatGPT), and tab-autocomplete code (GitHub Copilot running in Visual Studio Code).

References

[1] I. R. Ward, D. M. Asmar, M. Arief, J. K. Mike, and M. J. Kochenderfer, "Optimal control of mechanical ventilators with learned respiratory dynamics," in *International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2024, pp. 192–198.

[2] NVIDIA, "Cosmos world foundation model platform for physical AI," *arXiv preprint arXiv:2501.03575*, 2025.

[3] A. N. Angelopoulos and S. Bates, "Conformal Prediction: A Gentle Introduction," *Foundations and Trends in Machine Learning*, vol. 16, no. 4, pp. 494–591, 2023.

[4] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. Springer, 2009.

[5] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–58, 2009.

[6] M. Ho, A. Farid, and A. Majumdar, "Towards a framework for comparing the complexity of robotic tasks," in *International Workshop on the Algorithmic Foundations of Robotics*. Springer, 2022, pp. 273–293.

[7] J. L. Contreras, O. Shorinwa, and M. Schwager, "Safe, out-of-distribution-adaptive mpc with conformalized neural network ensembles," in *7th Annual Learning for Dynamics\& Control Conference*. PMLR, 2025, pp. 194–207.

[8] A. Vyas, N. Jammalamadaka, X. Zhu, D. Das, B. Kaul, and T. L. Willke, "Out-of-Distribution Detection Using an Ensemble of Self Supervised Leave-out Classifiers," in *European Conference on Computer Vision*, 2018, pp. 550–564.

[9] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight Uncertainty in Neural Network," in *International Conference on Machine Learning (ICML)*, 2015, pp. 1613–1622.

[10] J. Liu, Z. Lin, S. Padhy, D. Tran, T. Bedrax Weiss, and B. Lakshminarayanan, "Simple and Principled Uncertainty Estimation with Deterministic Deep Learning via Distance Awareness," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020, pp. 7498–7512.

[11] C. Richter and N. Roy, "Safe visual navigation via deep learning and novelty detection," in *Robotics: Science and Systems*, 2017.

[12] P. Oza and V. M. Patel, "C2ae: Class conditioned auto-encoder for open-set recognition," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2307–2316.

[13] K. Lee, K. Lee, H. Lee, and J. Shin, "A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[14] L. Bergman and Y. Hoshen, "Classification-Based Anomaly Detection for General Data," in *International Conference on Learning Representations (ICLR)*, 2020.

[15] J. Tack, S. Mo, J. Jeong, and J. Shin, "CSI: Novelty Detection via Contrastive Learning on Distributionally Shifted Instances," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020, pp. 11 839–11 852.

[16] Z. Xiao, Q. Yan, and Y. Amit, "Likelihood Regret: An Out-of-Distribution Detection Score For Variational Auto-encoder," in *Advances in Neural Information Processing Systems (NIPS)*, 2020.

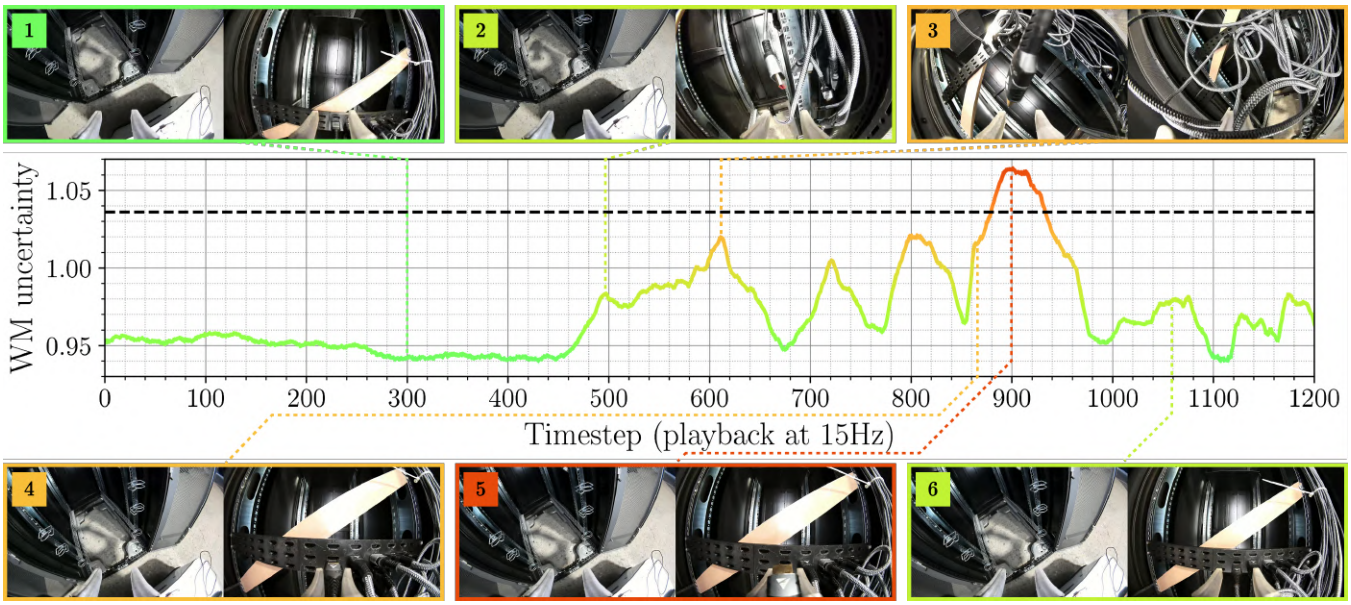


Fig. 7: An example failure detection on an unseen failure case, visualizing captures from the left and right gripper cameras. (1) Nominal behavior—the robot is not gripping or attempting to manipulate a cable. (2) The uncertainty rises as the gripper approaches a cable and begins a sequence of grab attempts. Uncertainty tends to increase when a cable head is in view. (3) The first of three aborted grab attempts, such periods are usually associated with increased uncertainty as grab attempts represent a small portion of the nominal training set. Note that these do not constitute failures. (4) Uncertainty increases rapidly as the robot focuses instead on unplugging a cable that is already plugged in (observe the force being applied onto the cable head by the compliant grippers). (5) An actual failure occurs (dropped cable) as the robot fumbles the cable head during the unplug maneuver. (6) With the dropped cable fully out of frame, the robot is again not gripping or attempting to manipulate a cable as in (1). Uncertainty again reduces to nominal levels. Note that the world model is trained on proprioceptive state data, and is action conditioned, so some part of the predicted uncertainty likely stems from proprioceptive features not visualized in this figure.

[17] Z. Wang, B. Dai, D. Wipf, and J. Zhu, "Further Analysis of Outlier Detection with Deep Generative Models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020, pp. 8982–8992.

[18] I. R. Ward, M. Paral, K. Riordan, and M. J. Kochenderfer, "Improving the resilience of quadrotors in underground environments by combining learning-based and safety controllers," *International Conference on Control, Decision and Information Technologies*, 2025.

[19] C. Xu, T. K. Nguyen, E. Dixon, C. Rodriguez, P. Miller, R. Lee, P. Shah, R. Ambrus, H. Nishimura, and M. Itkina, "Can we detect failures without failure data? uncertainty-aware runtime failure detection for imitation learning policies," in *Robotics: Science and Systems*, 2025.

[20] E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan, "Do Deep Generative Models Know What They Don't Know?" in *Conference on Robot Learning (CoRL)*, 2019.

[21] P. Kirichenko, P. Izmailov, and A. G. Wilson, "Why Normalizing Flows Fail to Detect Out-of-Distribution Data," 2020.

[22] G. Zhou, H. Pan, Y. Lecun, and L. Pinto, "Dino-wm: World models on pre-trained visual features enable zero-shot planning," in *International Conference on Machine Learning*. PMLR, 2025, pp. 79 115–79 135.

[23] K. Nakamura, L. Peters, and A. Bajcsy, "Generalizing safety beyond collision-avoidance via latent-space reachability analysis," *arXiv preprint arXiv:2502.00935*, 2025.

[24] Y. Wu, R. Tian, G. Swamy, and A. Bajcsy, "From foresight to forethought: Vlm-in-the-loop policy steering via latent alignment," *arXiv preprint arXiv:2502.01828*, 2025.

[25] M. Ho, M. F. Ginting, I. R. Ward, A. Reinke, M. J. Kochenderfer, A.-a. Agha-Mohammadi, and S. Omidshafiei, "World model failure classification and anomaly detection for autonomous inspection," *IEEE International Conference on Robotics and Automation*, 2026.

[26] J. Seo, K. Nakamura, and A. Bajcsy, "Uncertainty-aware latent safety filters for avoiding out-of-distribution failures," in *9th Annual Conference on Robot Learning*.

[27] "Semantic anomaly detection with large language models," *Autonomous Robots*, vol. 47, no. 8, pp. 1035–1055, 2023.

[28] A. Elhafsi, R. Sinha, C. Agia, E. Schmerling, I. A. Nesnas, and M. Pavone, "Semantic anomaly detection with large language models," *Autonomous Robots*, vol. 47, no. 8, pp. 1035–1055, 2023.

[29] G. G. Pihlgren, F. Sandin, and M. Liwicki, "Improving image autoencoder embeddings with perceptual loss," in *IEEE International Joint Conference on Neural Networks (IJCNN)*, 2020.

[30] N. Japkowicz, C. Myers, M. Gluck *et al.*, "A novelty detection approach to classification," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 1995.

[31] S. Balasubramanian, A. Melendez-Calderon, A. Roby-Brami, and E. Burdet, "On the analysis of movement smoothness," *Journal of Neuroengineering and Rehabilitation*, vol. 12, no. 1, p. 112, 2015.

[32] H. Liu, Y. Zhang, V. Betala, E. Zhang, J. Liu, C. Ding, and Y. Zhu, "Multi-task interactive robot fleet learning with visual world models," in *Conference on Robot Learning (CoRL)*, 2024.

[33] A. N. Angelopoulos and S. Bates, "Conformal prediction: A gentle introduction," *Foundations and Trends in Machine Learning*, vol. 16, no. 4, pp. 494–591, 2023.

[34] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," in *Robotics: Science and Systems*, 2023.

APPENDIX

All experiments were executed on a `gpu_1x_h100_pcie` instance on Lambda Cloud. The system was equipped with a single NVIDIA H100 PCIe GPU with 80 GB of VRAM (driver version 570.158.01, CUDA 12.8), paired with an Intel Xeon Platinum 8480+ CPU.