How Safe Am I Given What I See? Calibrated Prediction of Safety Chances for Image-Controlled Autonomy

Zhenjiang Mao Z.MAO@UFL.EDU
Carson Sobolewski CSOBOLEWSKI@UFL.EDU
Ivan Ruchkin IRUCHKIN@ECE.UFL.EDU
University of Florida

Editors: A. Abate, K. Margellos, A. Papachristodoulou

Abstract

End-to-end learning has emerged as a major paradigm for developing autonomous controllers. Unfortunately, with its performance and convenience comes an even greater challenge of safety assurance. A key factor in this challenge is the absence of low-dimensional and interpretable dynamical states, around which traditional assurance methods revolve. Focusing on the online safety prediction problem, this paper systematically investigates a flexible family of learning pipelines based on generative world models, which do not require low-dimensional states. To implement these pipelines, we overcome the challenges of missing safety labels under prediction-induced distribution shift and learning safety-informed latent representations. Moreover, we provide statistical calibration guarantees for our safety chance predictions based on conformal inference. An extensive evaluation of our predictor family on two image-controlled case studies, a racing car and a cart pole, delivers counterintuitive results and highlights open problems in deep safety prediction.

Keywords: pixel-to-action control, world models, confidence calibration, conformal prediction

1. Introduction

To handle real-world complexity, autonomous systems increasingly rely on high-resolution sensors like cameras/lidars and large learning architectures to process the sensing data. It has become increasingly commonplace to use end-to-end reinforcement and imitation learning Codevilla et al. (2018); Tomar et al. (2022); Betz et al. (2022), conveniently bypassing conventional intermediate steps such as state estimation and planning. While end-to-end controllers can learn complex behaviors from large raw datasets, they complicate safety assurance of such systems Fulton et al. (2019).

Traditional approaches to ensure safety (e.g., a car avoids an obstacle) rely on the notion of a *low-dimensional state* with a physical meaning (e.g., the car's position and velocity). For example, reachability analysis propagates the system's state forward Bansal et al. (2017); Chen and Sankaranarayanan (2022), barrier functions synthesize safe low-dimensional controllers Ames et al. (2019); Xiao et al. (2023), and trajectory predictors output future states of agents Salzmann et al. (2020); Teeti et al. (2022). These methods do not straightforwardly scale to image-based controllers with thousands of inputs without an exact physical interpretation. Instead, to take images into account, they would need abstractions of sensing/perception, which are either system-specific and effortful to build (e.g., modeling the ray geometry behind pixel values Santa Cruz and Shoukry (2022)) or simplified and potentially inaccurate (e.g., a linear overapproximation of vision Hsieh et al. (2022)).

The problem raised in this paper is the reliable online prediction of safety for an autonomous system with an image-based controller *without access to a physically meaningful low-dimensional dynamical state or model*. For instance, given an image, what is the probability that a racing car stays within the track's bounds in the next 5 seconds? Here, the reliability requirement, also known as *calibration* Guo et al. (2017), is to provide an upper bound on the difference between the estimated and true probabilities. Devoid of typical dynamical models, this setting calls for a careful combination of image-based prediction and statistical guarantees. Existing works focus on learning to control rather than assuring image-based systems Hafner et al. (2019), assume access to a true low-dimensional initial state Lindemann et al. (2023), or do not provide any reliability guarantees Acharya et al. (2022). Nonetheless, reliably solving the prediction problem would enable downstream safety interventions like human handoff or fallback control, which are outside of this paper's scope.

This paper systematically formalizes and investigates a *family of learning-based pipelines* for safety prediction. This family varies in its specificity to a given controller and in its modularity — from single-step supervised learning (which serves as a baseline) to learning intermediate representations, inspired by generative architectures for reinforcement learning known as *world models* Ha and Schmidhuber (2018).

To provide reliability/calibration guarantees, we combine *post-hoc calibration* Zhang et al. (2020) with the recently popular distribution-free technique of *conformal prediction* Vovk et al. (2005); Lei et al. (2018). This enables tuning the predictive safety chances orthogonally to the choice of the prediction pipeline and providing statistical bounds for them from validation data.

We perform extensive experiments on two popular benchmarks in the OpenAI Gym Brockman et al. (2016): racing car and cart pole. Focusing on long prediction horizons, we discover counterintuitive impacts of modularity and controller-specificity in our pipeline family. We also find that predicting well-calibrated safety chances is easier than safety labels over longer horizons. Due to space limits, we report only the key results and refer the reader to the full online version Mao et al. (2024).

Thus, we make three contributions: (i) an organization of learning pipelines for online safety prediction in image-driven autonomy; (ii) a conformal post-hoc calibration technique with statistical guarantees for safety chances; and (iii) an extensive evaluation of our predictors on two case studies.

After notation in Sec. 2, we describe our predictor family in Sec. 3 and conformal calibration in Sec. 4. Case study results are given in Sec. 5, then we review related work in Sec. 6 and conclude.

2. Notation and Problem Formulation

Definition 1 (Dynamical system) *A discrete-time* dynamical system $s = (\mathbf{X}, \mathbf{Z}, \mathbf{U}, h, f, o, x_0, \varphi)$ *consists of:* state space $\mathbf{X}$*, containing continuous states* x*;* observation space $\mathbf{Z}$*, containing images* y*;* action space $\mathbf{U}$*, with discrete/continuous commands* u*;* image-based controller $h : \mathbf{Z} \to \mathbf{U}$*, typically implemented by a neural network;* dynamical model $f : \mathbf{X} \times \mathbf{U} \to \mathbf{X}$*, which sets the next state from a past state and an action (unknown to us);* observation model $o : \mathbf{X} \to \mathbf{Z}$*, which generates an observation based on the state (unknown to us);* initial state x_0*, from which the system starts executing;* state-based safety property $\varphi : \mathbf{X} \to \{0, 1\}$*, which determines if a state* x *is safe.*

We focus on systems with observation spaces with thousands of pixels and unknown non-linear dynamical and observation models. In such systems, while the state space $\mathbf{X}$ may be conceptually known (if only to define φ), it is not necessary (and often difficult) to construct f and o because the controller acts directly on the observation space $\mathbf{Z}$. Without relying on f and o, end-to-end methods like deep reinforcement learning Mnih et al. (2015) and imitation learning Hussein et al. (2017) train controller h on the observation data. Once deployed in state x_0, the system executes a *trajectory*, which is a sequence $\{x_i, y_i, u_i\}_{i=0}^{t}$ up to time t, where: $x_{i+1} = f(x_i, u_i)$; $y_i = o(x_i)$; $u_i = h(y_i)$.

Instead of using this model, we extract predictive information from three sources of data. First, we will use the current observation, y_i, at some time i. For example, when a car is at the edge of the track, it has a higher probability of being unsafe in the next few steps. Second, past observations $y_{i-m+1}, \dots, y_i$ provide dynamically useful features only available in time series, such as the speed/direction of motion. For example, just before a car enters a turn, the sequence of past observations implicitly informs how hard it will be to remain safe. Third, not only do observations inform safety, but so do the controller's past/present outputs $h(y_{i-m+1}, \dots, y_i)$. For instance, if by mid-turn the controller has not changed the steering angle, it is less likely to navigate this turn safely.

Our goal is to predict the system's safety $\varphi(x_{i+k})$ at time $i+k$ given a series of m observations $\mathbf{y}_i = (y_{i-m+1}, ..., y_i)$. This sequence of observations does not, generally, determine the true state x_i (e.g., when $m = 1$, function o may not be invertible). This leads to *partial state observability*, which we model stochastically. Specifically, we say that from the predictor's perspective, x_i is drawn from some belief distribution $\mathcal{D}_{\mathbf{y}_i}$. This induces a distribution of subsequent trajectories and transforms future safety $\varphi(x_{i+k})$ into a Bernoulli random variable. Therefore, we will estimate the conditional probability $P(\varphi(x_{i+k}) \mid \mathbf{y}_i)$ and provide an error bound on our estimates. Note that process noise in f_d and measurement noise in f_o are orthogonal to the issue of partial observability; nonetheless, both of these noises are supported by our approach and would be treated as part of the stochastic uncertainty in the future φ. To sum up the above, we arrive at the following problem description.

Problem (Calibrated safety prediction) *Given horizon $k > 0$, confidence $\alpha \in (0, 0.5)$, and observations $\mathbf{y}_i$ from system s with unknown f and o,* estimate future safety chance $P(\varphi(x_{i+k}) \mid \mathbf{y}_i)$ *and* provide an upper bound *for the estimation error that holds in at least $1 - \alpha$ cases.*

To address this problem, we will build two types of predictors: for safety labels and chances.

Definition 2 (Safety label predictor) *For horizon $k > 0$, a* safety label predictor $\rho : \mathbf{Z}^m \to \{0, 1\}$ *predicts the safety outcome $\varphi(x_{i+k})$ at time $i + k$.*

Definition 3 (Safety chance predictor) *For horizon $k > 0$, a* safety chance predictor $g : \mathbf{Z}^m \to [0, 1]$ *predicts the safety chance $P(\varphi(x_{i+k}) \mid \mathbf{y}_i)$ at time $i + k$.*

Controller-specific (resp. *controller-independent*) predictors are trained on observation-controller (resp. observation-action) datasets. We expect that using controller-independent predictors trades off some prediction power for generalizability, and it is the first flexibility dimension of our family of learning pipelines — **controller-specificity**. We compare these two types of predictors in Sec. 5.

Our problem setting assumes that all our training data for training these predictors is obtained offline. We collect two types of datasets, with a conveniently unified notation $\mathbf{Z}$ and $\mathbf{z}$ to mean either of the dataset types. The first type in Def. 4 only contains data from a specific controller, while the second in Def. 5 can mix controllers due to explicit actions.

Definition 4 (Observation-controller dataset) *An* observation-controller dataset $\mathbf{Z} = \{(\mathbf{z}_j, \varphi_j) \mid j = 1, \dots, N\}$ *consists of pairs of m-long sequences $\mathbf{z}_j := (y_{i-m}, ..., y_i)$ for some time i and safety labels $\varphi_j := \varphi(x_{i+k})$ at k steps later, collected by executing a fixed controller h.*

Definition 5 (Observation-action dataset)

An observation-action dataset $\mathbf{Z} = \{(\mathbf{z}_j, \varphi_j) \mid j = 1, \dots, N\}$ *consists of pairs of m-long sequences of paired observations and corresponding actions $\mathbf{z}_j = ((y_{i-m+1}, u_{i-m+1}), ..., (y_i, u_i))$, and safety labels $\varphi_j := \varphi(x_{i+k})$ obtained k steps later.*

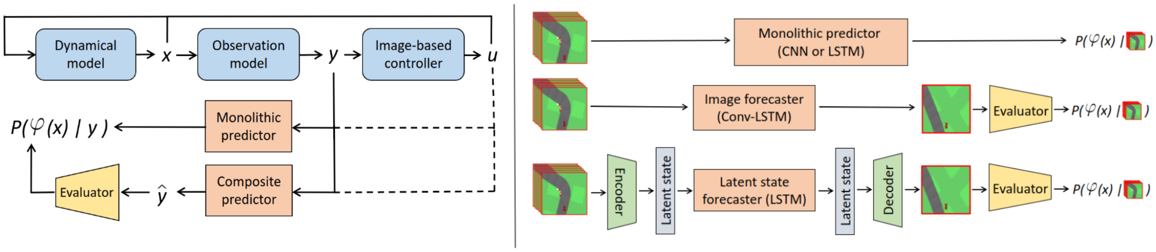

Figure 1: Left: Dynamical system with predictors. Arrows are dataflows, and dashes are controller dependencies. Right: monolithic, composite image-based, and latent-based predictors.

3. Safety Predictor Family

This section develops safety *label* predictors, and Sec. 4 transforms them into *chance* predictors.

3.1. Monolithic and Composite Label Predictors

The second dimension of our predictor family is **modularity**, which distinguishes three types of safety label predictors: monolithic predictors, composite image predictors, and composite latent predictors. The main distinction between monolithic and composite predictors is that the latter end with an *evaluator*, a separate component that determines the safety of the prediction — as shown in Fig. 1.

Monolithic predictors directly determine the future safety property based on the observations, as described in Def. 2. These predictors implement a baseline approach of single-step supervised binary classification similar to many existing papers, e.g., Strickland et al. (2018), by leveraging deep vision models (with convolutional layers) and time-series models (with recurrent layers). These models are trained on any training dataset $\mathbf{Z}_t$ with a fixed horizon k and a typical binary classification loss such as cross-entropy. The key drawback of monolithic predictors is their need to be completely re-trained to change the prediction horizon or other hyperparameters.

A composite predictor consists of multiple learning models. A *forecaster* model constructs the likely future observations, essentially approximating the dynamics f and observation model o. Another binary classifier — an *evaluator* — judges whether a forecasted observation is safe.

Definition 6 (Composite image predictor) *A* composite image predictor $v(f_g(\mathbf{z}_i))$ *consists of an* image forecaster $f_g : \mathbf{Z}^m \to \mathbf{Y}$*, which predicts the future observation* $\hat{y}_{i+k}$ *based on past* $\mathbf{z}_i$*; and an* evaluator $v : \mathbf{Y} \to \{0, 1\}$*, which determines the safety of the predicted images* $\varphi(\hat{y}_{i+k})$.

We implement the composite image predictor with a *convolutional LSTM* (conv-LSTM) Shi et al. (2015) for image forecasting and a *convolutional neural network* (CNN) for evaluating.

Definition 7 (Composite latent predictor) *A* composite latent predictor $v(d(f_l(e(\mathbf{z}_i))))$ *consists of an* autoencoder*, which provides an encoder* e *and decoder* d *over latent space* Θ *with state vectors* θ*; a* latent forecaster $f_l : \Theta^m \to \Theta$*, which predicts future latent state* $\hat{\theta}_{i+k}$ *based on the past* m *ones; and an* evaluator $v : \mathbf{Y} \to \{0, 1\}$*, which operates on the images from the decoder* d*.*

We implement composite latent predictors using a *Variational Autoencoder* (VAE) Kingma and Welling (2014), although we have experimented with vanilla autoencoders Hinton and Salakhutdinov (2006) and vector-quantized VAEs van den Oord et al. (2017) as well. Latent forecasting uses *Long-Short Term Memory* (LSTM) networks Lindemann et al. (2021), and evaluators are CNNs.

3.2. Training Process

Monolithic label predictors ρ evaluators v, and image forecasters f_g are trained with supervision on a training dataset $\mathbf{Z}_t$ of observation sequences "labeled" with safety or future images respectively. Latent forecasters f_l are trained on latent vector sequences, obtained from observations with an encoder e, which is trained as shown below. Monolithic predictors and evaluators use *cross entropy* (CE) loss. The *binary cross-entropy* (BCE) loss, i.e., average CE per pixel, is used for the conv-LSTM image forecaster f_g, and the *mean squared error* (MSE) loss is used for the LSTM latent forecaster f_l. We also use early stopping to reduce the learning rate as the loss drop slows.

For the VAE, the overall loss function $\mathcal{L}$ consists of three parts with regularizers λ_1, $\lambda_2 > 0$:

$$\mathcal{L} = \mathcal{L}_{recon} + \lambda_1 \mathcal{L}_{latent} + \lambda_2 \mathcal{L}_{safety}, \tag{1}$$

where the first is the *reconstruction loss* $\mathcal{L}_{recon}$, which quantifies how well an original image y is approximated by $d(e(y))$ using MSE. The second is the latent loss $\mathcal{L}_{latent}$, which uses Kullback-Leibler divergence to minimize the difference between two latent-vector probability distributions $e(y \mid \theta) \approx d(\theta \mid y)$, where θ is a latent vector and y is an input image. To preserve the information about safety in latent representations, we add a *safety loss* $\mathcal{L}_{safety}$, which is a BCE loss based on the true safety φ_i truth of the original images y_i and the safety evaluation of the reconstructed images:

$$\mathcal{L}_{safety} = \mathcal{L}_{CE}(v(d(e(y_i))), \varphi_i) \tag{2}$$

One challenge is that the balance of safety labels changes in $\mathbf{Z}_t$ with the horizon k, leading to imbalanced (less safe) data for higher horizons. To ensure a balanced label distribution, we resample with replacement for a 1:1 safe:unsafe class balance both in the training and testing datasets.

Another and greater challenge is the *distribution shift* between the original and forecasted images from the image/latent forecasters: the forecasted images are distorted (e.g., see the images in the online version Mao et al. (2024)) and they do not automatically come with safety labels because they are sampled from the forecasted space without a physical ground truth. However, these labels are needed to train high-performance evaluators on the forecasted images, as per Fig. 1.

To overcome the distribution shift, we implement two *specialized evaluators* to finish the last step that generates binary safety consequences from predicted images. One evaluator is trained with *data augmentation* by randomly adjusting the brightness, inverting, and adding Gaussian blur to the training image. The other evaluator implements *vision-based* logic with robust system-specific features. For example, for the racing car, we take advantage of the fixed location of the car in the image and the contrasting colors to determine whether the car is in a safe position. Specifically, we crop the image to the area directly surrounding the car and use the mean of the pixel values to determine whether the car is on or off the track. To overcome color inversion in image-based forecasters we use the median pixel value of the full image to determine whether the track surface is painted in an inverted, lighter-than-background color. We discuss the implications of such hand-tuning in Sec. 7.

4. Conformal Calibration for Chance Predictors

To turn a label predictor ρ into a chance predictor g, we perform post-hoc calibration over its normalized softmax scores Guo et al. (2017); Zhang et al. (2020). On a held-out *calibration dataset* $\mathbf{Z}_c$, we hyperparameter-tune the choice of a post-hoc calibrator: temperature scaling, logistic/beta calibration, histogram binning, isotonic regression, ensemble of near isotonic regression (ENIR),

Algorithm 1 Conformal calibration for chance predictions

```
**Input**: A validation dataset bin $B = \{b_k\}_{k=1,...}$ which each contains sequences of observations and safety $b_k = (y_k, \varphi(x_k))$, trained safety chance predictor $g$, and miscoverage level $\alpha$.
**Output**: confidence bound $c$ satisfying Eq. 3.
FUNCTION ConCali $(B, g, \alpha)$:
for $i = 1$ to $M$ do
    $B_i \leftarrow N$ i.i.d. samples from $B$ {obtain a resampled bin from original bin number $j$}
    $\overline{q}_i \leftarrow \frac{1}{N}\sum_{l=1}^{N} g(y_l),$ for each $y_l \in B_i$ {compute the mean safety chance prediction}
    $\overline{p_i} \leftarrow \frac{1}{N}\sum_{l=1}^{N} \varphi(x_l),$ for each $\varphi(x_l) \in B_i$ {compute the true safety chance}
    $\delta_i \leftarrow |\overline{q}_i - \overline{p_i}|$ {use the calibration error as a non-conformity score}
end for
$n \leftarrow \lceil (M/2+1)(1-\alpha) \rceil$ {compute the conformal quantile for the upper bound}
$c \leftarrow$ the $n$-th smallest value among $\delta_1, \ldots, \delta_M$ {pick the upper-bound value}
return $c$
```

and Bayesian binning into quantiles (BBQ). The one with the smallest calibration error gives us calibrated softmax values $g(y_i)$. To ensure sufficient samples in each bin, we perform *equal-frequency binning* as defined below to structure a *validation dataset* $\mathbf{Z}_v$, on which we obtain our guarantees.

Definition 8 (Equal-frequency binning) *Given a dataset* $\mathbf{Z}$, *split* $\mathbf{Z}$ *into* Q *bins* $\{B_j\}_{j=1}^{Q}$ *by a constant count of samples,* $\lfloor |\mathbf{Z}|/Q \rfloor$ *each. The resulting* $\{B_j\}_{j=1}^{Q}$ *is a* binned dataset.

From each bin B_j, we draw N i.i.d. samples with replacement M times to get *resampled bins* $\{B_{j1}, \ldots, B_{jM}\}$. For each B_{ji}, we calculate the average safety confidence score $\overline{g}_{ji}$, the true safety chance p_{ji} (i.e., the fraction of samples that are truly safe), and the calibration error $\delta_{ji} := |\overline{g}_{ji} - p_{ji}|$.

Given a bin number j, our goal is to build *prediction intervals* $[0, c_j]$ that contain the calibration error of the next (unknown) average confidence $\overline{g}_{j*}$ that falls into bin B_j with chance at least $1-\alpha$:

$$P(|g_{j*} - p_j| \leq c_j) \geq 1 - \alpha \text{ for each } j = 1 \ldots Q \tag{3}$$

Once we get a statistical upper bound c_j on the calibration error in bin j, it will be combined with a chance prediction g_{j*} into an *uncertainty-aware interval* $[g_{j*} - c_j; g_{j*} + c_j]$, which contains the true probability of safety in $1-\alpha$ cases. This guarantee is relative to the binned dataset fixed in Def. 8.

To give this guarantee, Alg. 1 applies *conformal prediction* Lei et al. (2018) by ranking chance errors in resampled bins and getting a quantile-based upper bound for a given bin. As in other works using conformal prediction Qin et al. (2021); Lindemann et al. (2023), this algorithm ensures Eq. 3. Note that bin averaging is necessary to obtain meaningful probabilities, rather than binary outcomes.

Theorem 9 (Theorem 2.1 in Lei et al. (2018)) *Given a dataset bin* $B = \{b_k\}_{k=1}^{K}$ *of i.i.d. state-observation pairs* $b_k = (x_k, y_k)$, *we obtain a collection of datasets* $\{B_j\}_{j=1}^{M}$ *by drawing* M *datasets of* N *i.i.d. samples from* B, *leading to datasets* B_j *to be drawn i.i.d. from a dataset distribution* $\mathcal{D}$. *Then for another, unseen dataset* $B_{M+1} \sim \mathcal{D}$, *safety chance predictor* g, *and miscoverage level* α, *calculating* $c = \texttt{ConCali}(B, g, \alpha)$ *leads to prediction intervals with guaranteed containment:*

$$P_{\mathcal{D}}(|\overline{q} - \overline{p}| \leq c) \geq 1 - \alpha,$$

where $\overline{q}$ *is the mean safety chance prediction score:* $\overline{q} \leftarrow \frac{1}{N}\sum_{l=1}^{N} g(y_l)$ *for each* $y_l \in B_{M+1}$, *and* $\overline{p}$ *is the mean true safety chance:* $\overline{p} \leftarrow \frac{1}{N}\sum_{l=1}^{N} \varphi(x_l)$ *for each safety label* $\varphi(x_l) \in B_{M+1}$.

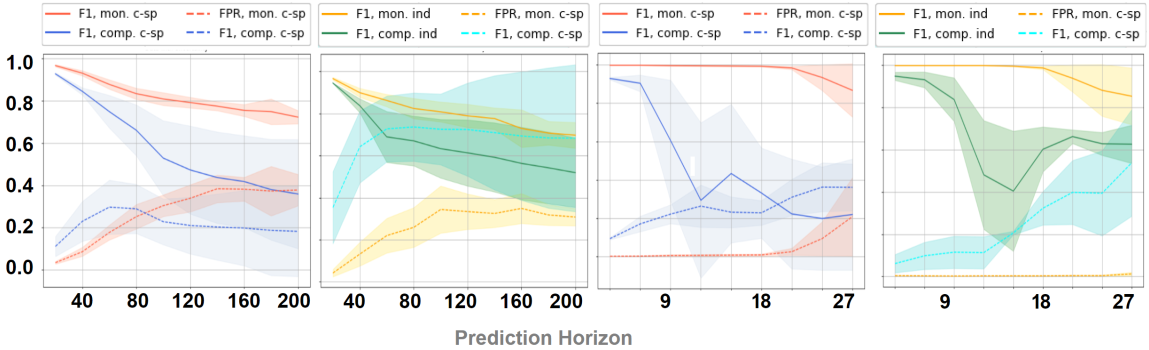

Figure 2: Performance of safety label predictors over varied horizons. L to R: (1) controller-specific ('c-sp') monolithic ('mon') vs. latent composite ('comp') for the racing car; (2) controller-independent ('ind') monolithic vs. latent composite for the racing car; (3) controller-specific monolithic vs. latent composite for the cart pole; (4) controller-independent monolithic vs. latent composite for the cart pole. Shaded uncertainty shows standard deviation due to different controllers and resampling.

5. Experimental Results

Systems. Our case studies are the *racing car* and *cart pole* from the OpenAI Gym Brockman et al. (2016), which we chose over autonomy datasets (e.g., Waymo Open) for two reasons. First, studying safety prediction requires a large dataset of *safety violations*, rarely found in real-world data. Second, simulated systems let us collect unbounded image data with direct access to the true state for evaluation. We defined the racing car's safety as staying within the track. The cart pole's safety is defined by the angles in the range of $[-6, 6]$ degrees (the full range is $[-48, 48]$ degrees).

Performance Metrics. *F1 score* is our main metric for evaluating safety label predictors, as it balances the precision and recall: $F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \times \frac{TP}{TP + \frac{FP+FN}{2}}$. *False Positive Rate* (FPR) is also a major concern in safety prediction (actually unsafe situations predicted as safe): FPR $= \frac{FP}{FP+TN}$. To evaluate our chance predictors, we compute the *Expected Calibration Error* (ECE, intuitively a weighted average difference between the predicted and true chances) and *Maximum Calibration Error* (MCE, intuitively the maximum above difference) Minderer et al. (2021).

Dataset. *Deep Q Networks* (DQN) implement three image-based controllers for each system. For racing car, we collected 240K samples for training and 60K for calibration/validation/testing each. For cart pole, we collected 90K samples for each of the four datasets. All images were processed with normalization and grayscale conversion. Each study's data is randomly partitioned into four datasets: (i) predictors learn on the training dataset $\mathbf{Z}_t$; (ii) hyperparameter tuning and calibrator fitting happen on the calibration dataset $\mathbf{Z}_c$; (iii) conformal calibration guarantees are made based on the validation dataset $\mathbf{Z}_v$; and (iv) the performance metrics are computed on the test dataset $\mathbf{Z}_e$.

Training details. We used PyTorch 1.13.1 with the Adam optimizer for training. The maximum training epoch is 500 for VAEs and 100 for predictors. The safety loss in Eq. 1 uses $\lambda_1 = 1$ and $\lambda_2 = 4096$, which equals the total pixel count in our images. The miscoverage level is $\alpha = 0.05$. The remaining hyperparameters are found in the online supplement Mao et al. (2024). We released our code at `https://github.com/maozj6/hsai-predictor`.

5.1. Comparative Results

Below we make comparisons and ablations along the four key dimensions of our predictor family.

Result 1: Monolithic predictors outperforms composite ones. Monolithic predictors do not learn the underlying dynamics, so we hypothesized that they would do well on short horizons but lose to composite predictors on longer horizons. To our surprise, as illustrated in the 1st and 3rd plot of Fig. 2, the performance of composite predictors degrades faster for longer horizons; the only aspect in which composite predictors excelled was a better FPR for the racing car, which may be desirable in safety-critical systems. We attribute the degradation of composite predictors to the challenge of learning coherent long-term latent dynamics, which remains an open problem.

Result 2: Latent predictors exceed the performance of image-based ones. Latent predictors outperform image-based ones both in F1 and FPR, for two reasons. The first is that the efficient compression by a safety-informed VAE supports generalizable learning of the dynamics: note the performance drop of image-based predictors when the horizon exceeds the training sequence length. Second, image-based forecasters tend to induce a stronger distribution shift on the forecasted images, hence disrupting the evaluator. This result is shown on Mao et al. (2024) due to the space limit.

Result 3: Controller-independent and controller-specific predictors show comparable performance. We hypothesized that controller-specific predictors would work better due to less variance. For composite predictors, the results are inconsistent with our hypothesis (see 1st vs 2nd, and 3rd vs 4th plots in Fig. 2). For monolithic predictors, we were surprised to see no significant difference between controller-specific and independent ones in F1 scores, while the independent ones do slightly better in FPR. This means that monolithic predictors obtain their versatility virtually “for free”.

Result 4: Calibrated predictors are superior to uncalibrated ones. An example reliability diagram (1st plot of Fig. 3) shows a trend we saw among uncalibrated predictors: underconfident for rejected class (below 0.5) and overconfident for the chosen class (above 0.5). Our calibration reduces the overconfidence to be within our conformal prediction intervals (2nd plot of Fig. 3) and leads to lower ECE/MCE even over long horizons (3rd plot of Fig. 3). The best calibrators were the isotonic regression (racing car) and ENIR (cart pole). Thus, we conclude that predicting the safety chance, rather than the label, is a more suitable formulation for highly uncertain image-controlled autonomy.

Result 5: Conformal calibration coverage is reliable. Supporting our theoretical claims, the predicted intervals for calibration errors contained the true error values from the test data. One example of these intervals is shown in the 4th plot of Fig. 3, where most error samples fall into the predicted intervals. Our average error bound is 0.022 for the racing car and 0.0041 for the cartpole. Therefore, our predicted chance ± its bin’s calibration bound can be used reliably and informatively online.

Additional results, illustrations and performance of different evaluators can be found in the extended online version: Mao et al. (2024).

6. Related Work

Performance and safety evaluation for autonomy. Recent research enables autonomous self-evaluation of competency Basich et al. (2022). Performance metrics vary significantly, including the time to goal navigation and safety violations. For systems with access to low-dimensional states, performance degradation has been measured with *hand-crafted indicators*, referred to as robot vitals Ramesh et al. (2022), alignment checkers Gautam et al. (2022), assumption monitors Ruchkin et al. (2022), and operator trust Conlon et al. (2022). These indicators rely on domain knowledge and careful offline analysis, which are difficult to obtain and perform for high-dimensional systems.

Autonomy with neural network (NN) controllers is difficult to analyze and vulnerable to distribution shift Moreno-Torres et al. (2012); Huang et al. (2020). Model-based *closed-loop verification approaches* analyze the safety of a NN-controlled system Ivanov et al. (2021); Tran et al. (2022),

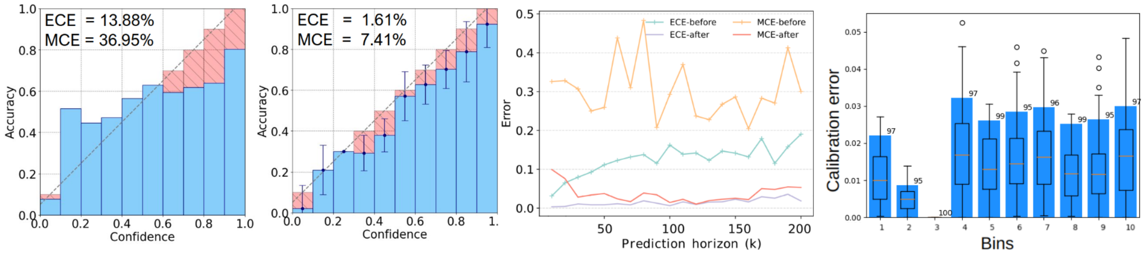

Figure 3: Calibration of a monolithic predictor for the racing car with horizon $k = 100$. L to R: (1) uncalibrated; (2): calibrated w/ isotonic regression, conformal bounds for $\alpha = 0.05$; (3) ECE/MCE over varied k before/after calibration; (4) our conformal bounds for $\alpha = 0.05$ (blue) contain more than 95% of the true calibration errors (box and whisker plots).

but for vision-based systems, require detailed modeling and have limited scalability Santa Cruz and Shoukry (2022); Hsieh et al. (2022). On the other hand, *model-free safety predictions* rely on the correlation between performance and uncertainty measures/anomaly scores, such as autoencoder reconstruction errors Stocco et al. (2020) and distances to training data Yang et al. (2022); however, such approaches fail to fully utilize the information about the system's closed-loop behavior. Striking the balance between model-based and model-free approaches are *black-box statistical methods* for risk assessment and safety verification Zarei et al. (2020); Michelmore et al. (2020); Qin et al. (2022); Cleaveland et al. (2022). They usually require low-dimensional states and rich outcome labels (e.g., signal robustness) — an assumption that our work relaxes while addressing a similar problem.

Trajectory prediction. A system's safety can be inferred from its *predicted trajectories*. Model-based approaches can estimate collision risks with, say, bicycle dynamics and Kalman filtering Ammoun and Nashashibi (2009); Lefèvre et al. (2014). One approach to safety guarantees is *Hamilton-Jacobi* (HJ) reachability Li et al. (2021); Nakamura and Bansal (2022), which requires model-based precomputation. Among many deep learning-based predictors Huang et al. (2022), a recently popular architecture, *Trajectron++* Salzmann et al. (2020), takes in high-dimensional scene graphs and outputs future agent trajectories. A conditional VAE Sohn et al. (2015) is used in Trajectron++ to add constraints at the decoding stage. Learning-based predictions can improve their reliability with conformal inference Lindemann et al. (2023); Muthali et al. (2023). In contrast, we eschew first-principles scene and state representations, instead using images and safety-informed latent vectors.

Safe control. Safe planning/control is a complementary problem to ours: controlling a system safely, usually with respect to a model, such as motion planning for uncertain systems Knuth et al. (2021); Hibbard et al. (2022); Chou and Tedrake (2023) and a growing body of work on control barrier functions Ames et al. (2019); Xiao et al. (2023). Some recent works are robust to errors of learning-based perception Dean et al. (2021); Yang et al. (2023), but they require low-dimensional states — a paradigm we aim to circumvent. On the other hand, our approach is constrained to a non-modifiable end-to-end controller, which can be implemented with the above methods.

Confidence calibration. Softmax scores of classification NNs tend to be miscalibrated as class probabilities Guo et al. (2017); Minderer et al. (2021), as measured by the *Brier score* and *Expected Calibration Error* (ECE). Calibration approaches can be *extrinsic (post-hoc)*, added on top of a trained network like Platt scaling Platt (1999), isotonic regression Zadrozny and Elkan (2002), and histogram/Bayesian binning Naeini et al. (2015) — and *intrinsic*, which modify the training, such as ensembles Zhang et al. (2020), adversarial training Lee et al. (2018), and learning from hints De-

Vries and Taylor (2018), error distances Xing et al. (2019), or true class probabilities Corbiere et al. (2019). Similar techniques have been used for regression NNs Vovk et al. (2020); Marx et al. (2022). To obtain calibration guarantees for safety chances, past works require a low-dimensional model-based setting Ruchkin et al. (2022); Cleaveland et al. (2023). To the authors' knowledge, such guarantees have not been instantiated for model-free autonomy.

Deep surrogate models. Trajectory predictions are challenging for systems with high-resolution images and complex dynamics. Physics-based methods and classical machine learning perform poorly in these settings, leading to the almost exclusive application of NNs. Vanilla deep *sequence prediction models*, originating in video prediction Oprea et al. (2022), are challenging to train because they require long horizons and many samples from each controller. However, we add observations and actions to sequence predictors to improve their performance and meantime propose different architectures in monolithic pipelines Strickland et al. (2018).

High-dimensional data contains redundant and irrelevant information, leading to higher computational costs. *Generative adversarial networks* (GANs) map observations into a low-dimensional latent space, enabling conventional assurance Katz et al. (2022). Low-dimensional representations can rely on conformal inference for performance monitoring Boursinos and Koutsoukos (2021). *World models* Ha and Schmidhuber (2018); Deng et al. (2023) provide surrogate training for controllers in reinforcement learning, instead of real data, showing high performance with the actor-critic *DreamerV2* algorithm Hafner et al. (2021). Such models implicitly learn the dynamics with recurrent NNs and use a VAE to learn the observation-latent space mapping. Our composite latent approach is inspired by recent adaptations of world models for trajectory prediction Acharya et al. (2023) and competency assessment Acharya et al. (2022). Lately, transformer-based world models Micheli et al. (2023); Robine et al. (2023) achieve more vivid image predictions. Our composite latent approach adopts the original world models with VAEs and recurrent NNs, which we compare with standalone recurrent/convolutional NNs; thus, our predictor family provides a systematic characterization of the latest prediction approaches from the literature. On top of these approaches, we have developed chance calibration guarantees, which were previously unexplored.

7. Discussion

Limitations: our predictor family's scope is limited to systems where safety can be inferred from raw sensor data. Also, our safety evaluators are system-specific, and developing them may require substantial effort; however, this effort is balanced with the savings of not designing state representations and dynamical models (both of which would be system-specific as well). Besides, obtaining negative safety samples can have a high cost in reality, which can be overcome with transfer learning from simulation and generative models.

Due to space limits, we did not report some poorly-performing techniques. For instance, *flexible-horizon predictors*, which predict the time until the next safety violation, failed to train due to a complex and insufficiently regularized prediction space. Also, *latent space evaluators* (as opposed to image evaluators, which we use) have shown particularly poor performance — providing evidence that latent vectors failed to preserve sufficient safety information. Thus, learning meaningful safety-informed representations for autonomy remains an open problem Liu et al. (2023).

Future work includes adding physical constraints to latent states as in neural ODEs Wen et al. (2022), jointly learning forecasters and evaluators to overcome distribution shifts, using transformer world models Micheli et al. (2023); Robine et al. (2023), and applications to physical systems.

Acknowledgments

The authors would like to thank Kaleb Smith, Yuang Geng, and anonymous reviewers for their insightful feedback on this research.

References

Aastha Acharya, Rebecca Russell, and Nisar R. Ahmed. Competency Assessment for Autonomous Agents using Deep Generative Models. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8211–8218, October 2022. doi: 10.1109/IROS47612.2022.9981991. ISSN: 2153-0866.

Aastha Acharya, Rebecca Russell, and Nisar R. Ahmed. Learning to Forecast Aleatoric and Epistemic Uncertainties over Long Horizon Trajectories. In *Proc. of ICRA 2023*, February 2023. doi: 10.48550/arXiv.2302.08669. arXiv:2302.08669 [cs].

Aaron D. Ames, Samuel Coogan, Magnus Egerstedt, Gennaro Notomista, Koushil Sreenath, and Paulo Tabuada. Control Barrier Functions: Theory and Applications. In *2019 18th European Control Conference (ECC)*, pages 3420–3431, June 2019. doi: 10.23919/ECC.2019.8796030.

Samer Ammoun and Fawzi Nashashibi. Real time trajectory prediction for collision risk estimation between vehicles. In *2009 IEEE 5th International Conference on Intelligent Computer Communication and Processing*, pages 417–422, 2009. doi: 10.1109/ICCP.2009.5284727.

Somil Bansal, Mo Chen, Sylvia Herbert, and Claire J. Tomlin. Hamilton-Jacobi reachability: A brief overview and recent advances. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 2242–2253, December 2017. doi: 10.1109/CDC.2017.8263977.

Connor Basich, Justin Svegliato, Kyle H. Wray, Stefan Witwicki, Joydeep Biswas, and Shlomo Zilberstein. Competence-Aware Systems. *Artificial Intelligence*, page 103844, December 2022. ISSN 0004-3702. doi: 10.1016/j.artint.2022.103844. URL https://www.sciencedirect.com/science/article/pii/S0004370222001849.

Johannes Betz, Hongrui Zheng, Alexander Liniger, Ugo Rosolia, Phillip Karle, Madhur Behl, Venkat Krovi, and Rahul Mangharam. Autonomous Vehicles on the Edge: A Survey on Autonomous Vehicle Racing. *IEEE Open Journal of Intelligent Transportation Systems*, 3:458–488, 2022. ISSN 2687-7813. doi: 10.1109/OJITS.2022.3181510. Conference Name: IEEE Open Journal of Intelligent Transportation Systems.

Dimitrios Boursinos and Xenofon Koutsoukos. Assurance monitoring of learning-enabled cyber-physical systems using inductive conformal prediction based on distance learning. *AI EDAM*, 35 (2):251–264, May 2021. ISSN 0890-0604, 1469-1760. Publisher: Cambridge University Press.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI Gym, June 2016. URL http://arxiv.org/abs/1606.01540. arXiv:1606.01540 [cs].

Xin Chen and Sriram Sankaranarayanan. Reachability Analysis for Cyber-Physical Systems: Are We There Yet? In *NASA Formal Methods Symposium*, pages 109–130. Springer, 2022.

Glen Chou and Russ Tedrake. Synthesizing Stable Reduced-Order Visuomotor Policies for Nonlinear Systems via Sums-of-Squares Optimization, September 2023. URL http://arxiv.org/abs/2304.12405. arXiv:2304.12405 [cs, eess, math].

Matthew Cleaveland, Lars Lindemann, Radoslav Ivanov, and George J. Pappas. Risk verification of stochastic systems with neural network controllers. *Artificial Intelligence*, 313:103782, December 2022. ISSN 0004-3702. doi: 10.1016/j.artint.2022.103782. URL https://www.sciencedirect.com/science/article/pii/S0004370222001229.

Matthew Cleaveland, Oleg Sokolsky, Insup Lee, and Ivan Ruchkin. Conservative Safety Monitors of Stochastic Dynamical Systems. In *Proc. of the NASA Formal Methods Conference*, May 2023.

Felipe Codevilla, Matthias Miiller, Antonio López, Vladlen Koltun, and Alexey Dosovitskiy. End-to-End Driving Via Conditional Imitation Learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–9, Brisbane, Australia, May 2018. IEEE Press. doi: 10.1109/ICRA.2018.8460487. URL https://doi.org/10.1109/ICRA.2018.8460487.

Nicholas Conlon, Daniel Szafir, and Nisar Ahmed. "I'm Confident This Will End Poorly": Robot Proficiency Self-Assessment in Human-Robot Teaming. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2127–2134, October 2022. doi: 10.1109/IROS47612.2022.9981653. ISSN: 2153-0866.

Charles Corbiere, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Perez. Addressing Failure Prediction by Learning Model Confidence. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/757f843a169cc678064d9530d12a1881-Abstract.html.

Sarah Dean, Andrew Taylor, Ryan Cosner, Benjamin Recht, and Aaron Ames. Guaranteeing Safety of Learned Perception Modules via Measurement-Robust Control Barrier Functions. In *Proceedings of the 2020 Conference on Robot Learning*, pages 654–670. PMLR, October 2021. URL https://proceedings.mlr.press/v155/dean21a.html. ISSN: 2640-3498.

Fei Deng, Junyeong Park, and Sungjin Ahn. Facing Off World Model Backbones: RNNs, Transformers, and S4. In *Proc. of NeurIPS 2023*, November 2023. doi: 10.48550/arXiv.2307.02064. arXiv:2307.02064 [cs].

Terrance DeVries and Graham W. Taylor. Learning Confidence for Out-of-Distribution Detection in Neural Networks, February 2018. URL http://arxiv.org/abs/1802.04865. arXiv:1802.04865 [cs, stat].

Nathan Fulton, Nathan Hunt, Nghia Hoang, and Subhro Das. Formal Verification of End-to-End Learning in Cyber-Physical Systems: Progress and Challenges. In *NeurIPS Workshop on Safety and Robustness in Decision Making*. arXiv, 2019. doi: 10.48550/arXiv.2006.09181. URL http://arxiv.org/abs/2006.09181. arXiv:2006.09181 [cs, stat].

Alvika Gautam, Tim Whiting, Xuan Cao, Michael A. Goodrich, and Jacob W. Crandall. A method for designing autonomous robots that know their limits. In *2022 International Conference* *on Robotics and Automation (ICRA)*, pages 121–127, 2022. doi: 10.1109/ICRA46639.2022.9812030.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 1321–1330, Sydney, NSW, Australia, August 2017. JMLR.org.

David Ha and Jürgen Schmidhuber. Recurrent World Models Facilitate Policy Evolution. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to Control: Learning Behaviors by Latent Imagination. In *In Proc. of International Conference on Learning Representations*, September 2019. URL https://openreview.net/forum?id=S1lOTC4tDS.

Danijar Hafner, Timothy P. Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering Atari with Discrete World Models. In *ICLR*, 2021. URL https://openreview.net/forum?id=0oabwyZbOu.

Michael Hibbard, Abraham P. Vinod, Jesse Quattrociocchi, and Ufuk Topcu. Safely: Safe Stochastic Motion Planning Under Constrained Sensing via Duality, March 2022. URL http://arxiv.org/abs/2203.02816. arXiv:2203.02816 [cs, eess].

G. E. Hinton and R. R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786):504–507, July 2006. doi: 10.1126/science.1127647. URL https://www.science.org/doi/10.1126/science.1127647. Publisher: American Association for the Advancement of Science.

Chiao Hsieh, Yangge Li, Dawei Sun, Keyur Joshi, Sasa Misailovic, and Sayan Mitra. Verifying Controllers With Vision-Based Perception Using Safe Approximate Abstractions. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 41(11):4205–4216, November 2022. ISSN 1937-4151. doi: 10.1109/TCAD.2022.3197508. Conference Name: IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems.

Xiaowei Huang, Daniel Kroening, Wenjie Ruan, James Sharp, Youcheng Sun, Emese Thamo, Min Wu, and Xinping Yi. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review*, 37:100270, 2020. ISSN 1574-0137. doi: https://doi.org/10.1016/j.cosrev.2020.100270. URL https://www.sciencedirect.com/science/article/pii/S1574013719302527.

Yanjun Huang, Jiatong Du, Ziru Yang, Zewei Zhou, Lin Zhang, and Hong Chen. A Survey on Trajectory-Prediction Methods for Autonomous Driving. *IEEE Transactions on Intelligent Vehicles*, 7(3):652–674, September 2022. ISSN 2379-8904. doi: 10.1109/TIV.2022.3167103. Conference Name: IEEE Transactions on Intelligent Vehicles.

Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. Imitation Learning: A Survey of Learning Methods. *ACM Computing Surveys*, 50(2):21:1–21:35, April 2017. ISSN 0360-0300. doi: 10.1145/3054912. URL https://doi.org/10.1145/3054912.

Radoslav Ivanov, Taylor Carpenter, James Weimer, Rajeev Alur, George Pappas, and Insup Lee. Verisig 2.0: Verification of Neural Network Controllers Using Taylor Model Preconditioning. In *Computer Aided Verification*, pages 249–262, Cham, 2021. Springer International Publishing. ISBN 978-3-030-81685-8.

Sydney M. Katz, Anthony L. Corso, Christopher A. Strong, and Mykel J. Kochenderfer. Verification of Image-Based Neural Network Controllers Using Generative Models. *Journal of Aerospace Information Systems*, 19(9):574–584, 2022. ISSN 1940-3151. doi: 10.2514/1.I011071. URL https://doi.org/10.2514/1.I011071. Publisher: American Institute of Aeronautics and Astronautics _eprint: https://doi.org/10.2514/1.I011071.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2014. URL https://openreview.net/forum?id=33X9fd2-9FyZd.

Craig Knuth, Glen Chou, Necmiye Ozay, and Dmitry Berenson. Planning With Learned Dynamics: Probabilistic Guarantees on Safety and Reachability via Lipschitz Constants. *IEEE Robotics and Automation Letters*, PP:1–1, March 2021. doi: 10.1109/LRA.2021.3068889.

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples, 2018.

Stéphanie Lefèvre, Dizan Vasquez, and Christian Laugier. A survey on motion prediction and risk assessment for intelligent vehicles. *ROBOMECH Journal*, 1(1):1, July 2014. ISSN 2197-4225. doi: 10.1186/s40648-014-0001-z. URL https://doi.org/10.1186/s40648-014-0001-z.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-Free Predictive Inference For Regression. *Journal of the American Statistical Association*, 2018. arXiv: 1604.04173.

Anjian Li, Liting Sun, Wei Zhan, Masayoshi Tomizuka, and Mo Chen. Prediction-Based Reachability for Collision Avoidance in Autonomous Driving. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7908–7914, May 2021. doi: 10.1109/ICRA48506.2021.9560790. ISSN: 2577-087X.

Benjamin Lindemann, Timo Müller, Hannes Vietz, Nasser Jazdi, and Michael Weyrich. A survey on long short-term memory networks for time series prediction. *Procedia CIRP*, 99:650–655, January 2021. ISSN 2212-8271. doi: 10.1016/j.procir.2021.03.088.

Lars Lindemann, Xin Qin, Jyotirmoy V. Deshmukh, and George J. Pappas. Conformal Prediction for STL Runtime Verification. In *Proc. of ICCPS'23*, San Antonio, TX, March 2023. arXiv. doi: 10.48550/arXiv.2211.01539. URL http://arxiv.org/abs/2211.01539. arXiv:2211.01539 [cs, eess].

Tian Yu Liu, Matthew Trager, Alessandro Achille, Pramuditha Perera, Luca Zancato, and Stefano Soatto. Meaning representations from trajectories in autoregressive models. *arXiv preprint arXiv:2310.18348*, 2023.

Zhenjiang Mao, Carson Sobolewski, and Ivan Ruchkin. How Safe Am I Given What I See? Calibrated Prediction of Safety Chances for Image-Controlled Autonomy, 2024. URL http://arxiv.org/abs/2308.12252. arXiv:2308.12252 [cs].

Charles Marx, Shengjia Zhao, Willie Neiswanger, and Stefano Ermon. Modular Conformal Calibration. In *Proceedings of the 39th International Conference on Machine Learning*, pages 15180–15195. PMLR, June 2022. URL https://proceedings.mlr.press/v162/marx22a.html. ISSN: 2640-3498.

Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=vhFu1Acb0xb.

Rhiannon Michelmore, Matthew Wicker, L. Laurenti, L. Cardelli, Y. Gal, and Marta Kwiatkowska. Uncertainty Quantification with Statistical Guarantees in End-to-End Autonomous Driving Control. *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020. doi: 10.1109/ICRA40945.2020.9196844.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the Calibration of Modern Neural Networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 15682–15694. Curran Associates, Inc., 2021.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Jose G. Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530, 2012. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2011.06.019. URL https://www.sciencedirect.com/science/article/pii/S0031320311002901.

Anish Muthali, Haotian Shen, Sampada Deglurkar, Michael H. Lim, Rebecca Roelofs, Aleksandra Faust, and Claire Tomlin. Multi-Agent Reachability Calibration with Conformal Prediction, April 2023. URL http://arxiv.org/abs/2304.00432. arXiv:2304.00432 [cs, eess].

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining Well Calibrated Probabilities Using Bayesian Binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015:2901–2907, January 2015. ISSN 2159-5399.

Kensuke Nakamura and Somil Bansal. Online Update of Safety Assurances Using Confidence-Based Predictions, December 2022. URL http://arxiv.org/abs/2210.01199. arXiv:2210.01199 [cs].

Sergiu Oprea, Pablo Martinez-Gonzalez, Alberto Garcia-Garcia, John Alejandro Castro-Vargas, Sergio Orts-Escolano, Jose Garcia-Rodriguez, and Antonis Argyros. A review on deep learning techniques for video prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2806–2826, jun 2022. doi: 10.1109/tpami.2020.3045007. URL https://doi.org/10.1109%2Ftpami.2020.3045007.

John C. Platt. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

Xin Qin, Yuan Xian, Aditya Zutshi, Chuchu Fan, and Jyotirmoy Deshmukh. Statistical Verification of Autonomous Systems using Surrogate Models and Conformal Inference. In *Proc. of ICCPS'22*, July 2021. URL http://arxiv.org/abs/2004.00279. arXiv: 2004.00279.

Xin Qin, Yuan Xia, Aditya Zutshi, Chuchu Fan, and Jyotirmoy V. Deshmukh. Statistical verification of cyber-physical systems using surrogate models and conformal inference. In *2022 ACM/IEEE 13th International Conference on Cyber-Physical Systems (ICCPS)*, pages 116–126, 2022. doi: 10.1109/ICCPS54341.2022.00017.

Aniketh Ramesh, Rustam Stolkin, and Manolis Chiou. Robot vitals and robot health: Towards systematically quantifying runtime performance degradation in robots under adverse conditions. *IEEE Robotics and Automation Letters*, 7(4):10729–10736, 2022. doi: 10.1109/LRA.2022.3192612.

Jan Robine, Marc Höftmann, Tobias Uelwer, and Stefan Harmeling. Transformer-based world models are happy with 100k interactions. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=TdBaDGCpjly.

Ivan Ruchkin, Matthew Cleaveland, Radoslav Ivanov, Pengyuan Lu, Taylor Carpenter, Oleg Sokolsky, and Insup Lee. Confidence Composition for Monitors of Verification Assumptions. In *ACM/IEEE 13th Intl. Conf. on Cyber-Physical Systems (ICCPS)*, pages 1–12, May 2022. doi: 10.1109/ICCPS54341.2022.00007.

Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-Feasible Trajectory Forecasting with Heterogeneous Data. In *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pages 683–700, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58523-5. doi: 10.1007/978-3-030-58523-5_40.

Ulices Santa Cruz and Yasser Shoukry. NNLander-VeriF: A Neural Network Formal Verification Framework for Vision-Based Autonomous Aircraft Landing. In Jyotirmoy V. Deshmukh, Klaus Havelund, and Ivan Perez, editors, *NASA Formal Methods*, Lecture Notes in Computer Science, pages 213–230, Cham, 2022. Springer International Publishing. ISBN 978-3-031-06773-0. doi: 10.1007/978-3-031-06773-0_11.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://papers.nips.cc/paper_files/paper/2015/hash/07563a3fe3bbe7e3ba84431ad9d055af-Abstract.html.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/8d55a249e6baa5c06772297520da2051-Paper.pdf.

Andrea Stocco, Michael Weiss, Marco Calzana, and Paolo Tonella. Misbehaviour prediction for autonomous driving systems. In *2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE)*, pages 359–371, 2020.

Mark Strickland, Georgios Fainekos, and Heni Ben Amor. Deep Predictive Models for Collision Risk Assessment in Autonomous Driving. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4685–4692, May 2018. doi: 10.1109/ICRA.2018.8461160. ISSN: 2577-087X.

Izzeddin Teeti, Salman Khan, Ajmal Shahbaz, Andrew Bradley, and Fabio Cuzzolin. Vision-based intention and trajectory prediction in autonomous vehicles: A survey. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, Lud De Raedt, Ed*, volume 7, pages 5630–5637, 2022.

Manan Tomar, Utkarsh Aashu Mishra, Amy Zhang, and Matthew E. Taylor. Learning Representations for Pixel-based Control: What Matters and Why? *Transactions on Machine Learning Research*, November 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=wIXHG8LZ2w.

Hoang-Dung Tran, Weiming Xiang, and Taylor T. Johnson. Verification Approaches for Learning-Enabled Autonomous Cyber–Physical Systems. *IEEE Design & Test*, 39(1):24–34, February 2022. ISSN 2168-2364. doi: 10.1109/MDAT.2020.3015712. Conference Name: IEEE Design & Test.

Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural Discrete Representation Learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/7a98af17e63a0ac09ce2e96d03992fbc-Abstract.html.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005 edition edition, March 2005. ISBN 978-0-387-00152-4.

Vladimir Vovk, Ivan Petej, Paolo Toccaceli, Alexander Gammerman, Ernst Ahlberg, and Lars Carlsson. Conformal calibrators. In *Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications*, pages 84–99. PMLR, August 2020. URL https://proceedings.mlr.press/v128/vovk20a.html. ISSN: 2640-3498.

Song Wen, Hao Wang, and Dimitris Metaxas. Social ODE: Multi-agent Trajectory Forecasting with Neural Ordinary Differential Equations. In *Computer Vision – ECCV 2022*, Lecture Notes in Computer Science, pages 217–233, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-20047-2. doi: 10.1007/978-3-031-20047-2_13.

Wei Xiao, Christos G. Cassandras, and Calin Belta. *Safe Autonomy with Control Barrier Functions: Theory and Applications*. Synthesis Lectures on Computer Science. Springer International Publishing, Cham, 2023. ISBN 978-3-031-27575-3 978-3-031-27576-0. doi: 10.1007/978-3-031-27576-0. URL https://link.springer.com/10.1007/978-3-031-27576-0.

Chen Xing, Sercan Arik, Zizhao Zhang, and Tomas Pfister. Distance-Based Learning from Errors for Confidence Calibration. In *In Proc. of International Conference on Learning Representations*, September 2019. URL https://openreview.net/forum?id=BJeB5hVtvB.

Shuo Yang, George J. Pappas, Rahul Mangharam, and Lars Lindemann. Safe Perception-Based Control under Stochastic Sensor Uncertainty using Conformal Prediction. In *CDC 2023*. arXiv, August 2023. arXiv:2304.00194 [cs, eess].

Yahan Yang, Ramneet Kaur, Souradeep Dutta, and Insup Lee. Interpretable Detection of Distribution Shifts in Learning Enabled Cyber-Physical Systems. In *2022 ACM/IEEE 13th International Conference on Cyber-Physical Systems (ICCPS)*, pages 225–235, May 2022. doi: 10.1109/ICCPS54341.2022.00027.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 694–699, New York, NY, USA, July 2002. Association for Computing Machinery. ISBN 978-1-58113-567-1. doi: 10.1145/775047.775151. URL https://dl.acm.org/doi/10.1145/775047.775151.

Mojtaba Zarei, Yu Wang, and Miroslav Pajic. Statistical verification of learning-based cyber-physical systems. In *Proceedings of the 23rd International Conference on Hybrid Systems: Computation and Control*, HSCC '20, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370189. doi: 10.1145/3365365.3382209. URL https://doi.org/10.1145/3365365.3382209.

Jize Zhang, Bhavya Kailkhura, and T. Yong-Jin Han. Mix-n-Match: ensemble and compositional methods for uncertainty calibration in deep learning. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20, pages 11117–11128. JMLR.org, July 2020.