NOISE-INFORMED CONFORMAL PREDICTION FOR TRAJECTORY WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Trajectory world models are predictive models that learn complex dynamics of diverse robotic systems from low-level states. Quantifying the uncertainty of trajectory world models is critical to ensuring the safety and reliability of downstream control and planning tasks. Conformal prediction (CP) is a model-agnostic uncertainty quantification framework with formal calibration guarantees, making it widely applicable to various learning-based dynamics models. However, trajectory data collected from robotic systems are inevitably affected by unbounded and temporally correlated observation noise, leading to an underestimation of the model's true prediction error. In such scenarios, existing methods cannot provide calibrated prediction sets for the system's true future states. To address this challenge, we propose Noise-Informed Conformal Prediction (NICP), which constructs an auxiliary covariance matrix by applying a covariance mapping to the noise information provided by the state estimator and then draws samples according to it to augment the observed prediction errors, thereby mitigating the underestimation of the model's true prediction error. Moreover, under certain conditions, we prove that the proposed method provides prediction sets with finite-sample marginal coverage guarantees. Experiments across different trajectory world models and both simulated and real-world robotic tasks show that our method can provide calibrated uncertainty estimates under noisy observations, while existing CP methods suffer from significant underestimation or overestimation in this setting.

1 INTRODUCTION

Trajectory world models (Yin et al., 2025; Wang et al., 2026) are learning-based predictive models that capture the complex dynamics of robotic systems, thereby enabling effective planning and control (Hansen et al., 2022; Schubert et al., 2023). However, when deployed in real-world environments, prediction errors can pose substantial risks to the safety and reliability of downstream planning and control (Dixit et al., 2023; Wei et al., 2024). Therefore, it is crucial to obtain well-calibrated estimates of the model's potential prediction errors, a task commonly referred to as uncertainty quantification (Kuleshov et al., 2018).

To quantify the uncertainty of trajectory world models, some recent studies (Seo et al., 2025; Li et al., 2025) have focused on ensemble-based methods for quantifying epistemic uncertainty, with the aim of ensuring the safety of control systems under out-of-distribution scenarios. Other studies (Frauenknecht et al., 2025; Aghabozorgi et al., 2026; Wan et al., 2026) further account for aleatoric uncertainty in model outputs to quantify the total uncertainty. Although these methods can provide explicit indicators of model uncertainty for downstream control tasks, they have *two major limitations. First,* they require tailored modifications to the model architecture or training procedure and therefore cannot be directly applied to a broad range of established trajectory world models. *Second,* their outputs are primarily indicative and lack formal calibration guarantees.

Given these limitations, many studies (Dixit et al., 2023; Wei et al., 2024; Nath et al., 2026; Chee et al., 2023) have employed conformal prediction, a model-agnostic framework, to quantify uncertainty in learned dynamics models. However, trajectory data collected from robotic systems are inevitably corrupted by observation noise, which typically arises from estimation errors of the state estimator (Hartley et al., 2020; Chen et al., 2023). Along a trajectory, observation noise may accumulate over time and become strongly correlated across time steps, leading to an underestimation

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

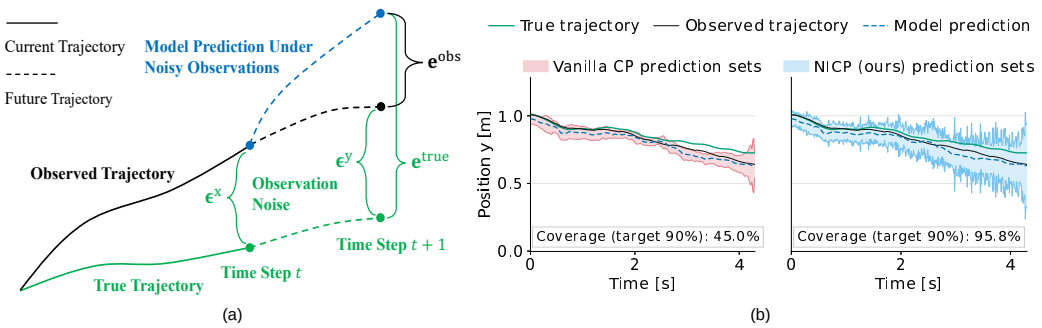


Figure 1: (a) Illustration of noisy observations in robotic systems, where ϵ^x and ϵ^y denote the observation noise at time steps t and $t+1$, respectively. **Green** denotes unobservable quantities, **Black** denotes observable quantities, and **Blue** denotes model predictions. $\mathbf{e}^{\text{obs}}$ and $\mathbf{e}^{\text{true}}$ are defined in Equation (10) and Equation (11), respectively. (b) Conformal prediction results for the y-axis position at each time step along a real-world Unitree Go2 trajectory, with the TDM trajectory world model (Schubert et al., 2023) used for one-step prediction.

of the model's true prediction error. Moreover, ground-truth trajectories are generally unavailable in real-world tasks, meaning that we have no data to calibrate the system's true prediction errors. In such scenarios, existing methods fail to provide calibrated prediction sets for the system's true future states, as illustrated in Figure 1.

To address this challenge, we propose a novel Noise-Informed Conformal Prediction (NICP) method, which can remain valid when all observations are corrupted by unbounded and correlated noise. NICP first constructs an auxiliary covariance matrix by applying a chosen covariance mapping to the noise information provided by the state estimator, and then draws samples according to the resulting covariance matrix to augment the observed prediction errors, thereby mitigating the underestimation of the true prediction errors. Moreover, under certain conditions, we prove that the proposed method provides finite-sample marginal coverage guarantees for the model's true prediction errors, and equivalently for the robotic system's true future states. Finally, we conduct experiments across multiple trajectory world models on a range of simulated and real-world robotic tasks. The experimental results show that, when the state estimator is reasonably consistent, our method can provide calibrated uncertainty estimates for trajectory world models under noisy observations, while directly applying existing conformal prediction methods to this scenario leads to substantial underestimation or overestimation.

Our main contributions include: **i)** We propose a novel conformal prediction method, named Noise-Informed Conformal Prediction (NICP), which can remain valid when *all observations are corrupted by unbounded and correlated noise*. **ii)** We theoretically prove that, under certain conditions, the prediction sets constructed by our method achieve finite-sample marginal coverage of the robotic system's true future states. **iii)** Experiments on a range of simulated and real-world robotic tasks demonstrate that, when the state estimator is reasonably consistent, NICP can provide calibrated uncertainty quantification for different trajectory world models under noisy observations.

2 Related Work

Uncertainty Quantification in Dynamics Learning. Uncertainty in learning-based models is commonly divided into two categories: aleatoric (data) uncertainty and epistemic (model) uncertainty (Abdar et al., 2021). Recent studies (Seo et al., 2025; Li et al., 2025) increasingly explored ensemble-based methods for quantifying the epistemic uncertainty of trajectory world models, while several works (Frauenknecht et al., 2025; Aghabozorgi et al., 2026; Wan et al., 2026) additionally accounted for aleatoric uncertainty in model outputs to quantify the total uncertainty. However, these methods typically require tailored modifications to the model architecture or training procedure, and the resulting uncertainty estimates generally lack formal guarantees. Meanwhile, conformal prediction, as a post-processing approach with formal coverage guarantees, has been widely used for uncertainty quantification in dynamics learning. Several works (Wei et al., 2024; Nath et al., 2026) directly applied split CP to calibrate the prediction errors of learned dynamics models. Other

works (Yang et al., 2023; Kulkarni et al., 2026) employed CP to address uncertainty induced by sensor noise. In addition, calibration failures arising from temporal dependence in trajectory data have also been widely studied (Dixit et al., 2023; Chee et al., 2023). However, these methods all require access to the ground-truth values of the quantities being calibrated, and therefore cannot provide finite-sample coverage guarantees for the system's true future states when no ground-truth states are available for calibration.

Conformal Prediction under Noisy Data. Trajectory data collected from robotic systems are often corrupted by observation noise, which typically arises from estimation errors of the state estimator (Hartley et al., 2020; Chen et al., 2023). Therefore, such noise generally follows a certain distribution and exhibits temporal correlations, leading to an underestimation of the model's true prediction error. Some CP methods (Gendler et al., 2021; Ghosh et al., 2023; Jeary et al., 2024) focused on restoring valid conformal coverage when the data were subject to bounded perturbations. Other studies (Einbinder et al., 2024; Penso & Goldberger, 2024; Cohen et al., 2025) found that when label noise was independent of the input features, vanilla conformal prediction tended to produce overly conservative prediction sets, motivating a reduction in set size. In addition, some studies (Einbinder et al., 2024; Feldman et al., 2026) proposed methods to correct the underestimation of vanilla conformal prediction under specific noise settings, but they required access to a subset of clean data or information that is unavailable in robotic applications. In the setting considered in this paper, however, all data are corrupted by unbounded noise, and the label noise is strongly correlated with the noise in the model inputs. Consequently, the above methods are not applicable to this setting and struggle to provide well-calibrated prediction sets.

3 PRELIMINARIES

Trajectory world model. For a robotic system, a trajectory world model (Yin et al., 2025) is a learning-based predictive model that estimates the system's future states $\mathbf{s}_{t+1}$ based on its current state $\mathbf{s}_t$ and action $\mathbf{a}_t$ at time step t. Mathematically, a trajectory world model characterizes a transition distribution:

$$p_{\boldsymbol{\theta}}(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t), \qquad \mathbf{s}_t \in \mathcal{S}, \quad \mathbf{a}_t \in \mathcal{A}, \tag{1}$$

where $\boldsymbol{\theta}$ denotes the model parameters, and $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces of the robotic system, respectively.

In practice, we typically use the mean of the output distribution as the prediction of the system's future state, which can then be used by downstream control tasks (Wang et al., 2026). In this case, the trajectory world model can be viewed as a predictor

$$f_{\boldsymbol{\theta}} : \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}, \qquad f_{\boldsymbol{\theta}}\big(\mathbf{s}_t, \mathbf{a}_t\big) = \widehat{\mathbf{s}}_{t+1}, \tag{2}$$

where $\widehat{\mathbf{s}}_{t+1}$ denotes the model prediction of the future state, and d_x and d_y denote the dimensions of the model input and output vectors, respectively. The model prediction error is then jointly affected by epistemic uncertainty arising from insufficient model training and aleatoric uncertainty inherent in the system state.

Conformal prediction. Conformal prediction is a model-agnostic uncertainty quantification method that returns a prediction set that is guaranteed to contain the ground truth with probability of at least $1 - \alpha$, given a user-specified miscoverage level α (Angelopoulos & Bates, 2023).

Specifically, let f be a fixed predictor and $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n+1}$ be exchangeable pairs, a weaker condition than being independent and identically distributed (i.i.d.). Split conformal prediction (Lei et al., 2018) computes

$$\mathrm{r}_i := s\big(f(\mathbf{x}_i) - \mathbf{y}_i\big), \qquad i = 1, \ldots, n, \tag{3}$$

where $s : \mathbb{R}^{d_y} \to [0, \infty)$ is a nonconformity score function. It then sets $k := \lceil (n+1)(1-\alpha) \rceil$, and defines $\widehat{q}_{1-\alpha}$ as the k-th order statistic of $\{\mathrm{r}_i\}_{i=1}^{n}$, with the convention $\widehat{q}_{1-\alpha} = +\infty$ if $k = n + 1$. By exchangeability,

$$\mathbb{P}\big(s\big(f(\mathbf{x}_{n+1}) - \mathbf{y}_{n+1}\big) \le \widehat{q}_{1-\alpha}\big) \ge 1 - \alpha, \tag{4}$$

which provides a calibrated characterization of the model prediction error at $\mathbf{x}_{n+1}$. Equivalently, this induces a calibrated prediction set for the unknown $\mathbf{y}_{n+1}$:

$$\mathbb{P}\big(\mathbf{y}_{n+1} \in \mathcal{C}_{1-\alpha}(\mathbf{x}_{n+1})\big) \ge 1 - \alpha, \tag{5}$$

where $\mathcal{C}_{1-\alpha}(\mathbf{x}_{n+1}) := \big\{\boldsymbol{y} : s\big(f(\mathbf{x}_{n+1}) - \boldsymbol{y}\big) \le \widehat{q}_{1-\alpha}\big\}$.

4 METHODOLOGY

4.1 PROBLEM STATEMENT

For a robotic task, suppose we have a pre-trained trajectory world model f_M and have additionally observed N task trajectories $\tau_1, \ldots, \tau_N$ outside the model-training dataset. When the robot starts a new task trajectory τ_{N+1}, at time step t, our goal is to provide a calibrated prediction set for the true future state of the robotic system.

Let $\mathbf{x}_i(t)$ and $\mathbf{y}_i(t)$, $i = 1, \ldots, N+1$ denote the true physical quantities corresponding to the input and output of f_M at time step t along trajectory τ_i. Due to estimation errors of the state estimator in the robotic system, these true physical quantities cannot be directly observed in practice. Instead, we only have access to noisy observations:

$$\mathbf{x}_i^{\text{obs}}(t) = \mathbf{x}_i(t) + \epsilon_i^x(t), \quad \mathbf{y}_i^{\text{obs}}(t) = \mathbf{y}_i(t) + \epsilon_i^y(t), \tag{6}$$

where $\epsilon_i^x(t)$ and $\epsilon_i^y(t)$ denote the observation noise of the robotic system states at time step t and $t+1$, respectively.

At the same time, many state estimators also provide uncertainty information in the form of estimated covariance matrices of their estimation errors (Hartley et al., 2020; Chen et al., 2023). Here, we use $\mathbf{\Sigma}_i^{x,\text{obs}}(t)$ to denote the covariance estimate returned by the estimator for $\epsilon_i^x(t)$. Many studies focus on constructing consistent state estimators (Huai et al., 2021; Lisus et al., 2023; Chen et al., 2024), meaning that their estimation errors are zero-mean and have covariance matrices equal to those reported by the estimators (Bar-Shalom et al., 2001). Mathematically,

$$\mathbb{E}[\epsilon_i^x(t) \mid \mathcal{I}_i(t)] = 0, \qquad \mathbb{E}\big[\epsilon_i^x(\epsilon_i^x)^\top \mid \mathcal{I}_i(t)\big] = \mathbf{\Sigma}_i^{x,\text{obs}}(t), \tag{7}$$

where $\mathcal{I}_i(t)$ denotes the information available at time step t along trajectory τ_i. In this case, the reported covariance matrices provide valuable information that can be directly exploited, and our study is built on this background. Specifically, noting that $\epsilon_i^y(t)$ can be viewed as $\epsilon_i^x(t+1)$, we make the following assumption:

Assumption 1. *Across different trajectories, the tuples*

$$(\mathbf{x}_i, \mathbf{y}_i, \epsilon_i^x, \epsilon_i^y, \mathbf{\Sigma}_i^{x,\text{obs}}), \qquad i = 1, \ldots, N+1, \tag{8}$$

are i.i.d. and the state estimator is conservative in expectation:

$$\mathbb{E}\big[\epsilon_i^x\big] = \mathbf{0}, \quad \mathbb{E}\big[\epsilon_i^y\big] = \mathbf{0}, \quad \mathbb{E}\big[\mathbf{\Sigma}_i^{x,\text{obs}}\big] \succeq \text{Cov}(\epsilon_i^x). \tag{9}$$

To simplify notation, we omit the time index and consider an arbitrary fixed time step. All results therefore apply pointwise to each time step along a trajectory. Additionally, all expectations and covariance matrices appearing in this paper are assumed to exist and be finite. Equation (9) imposes a substantially weaker condition than Equation (7): it can be viewed as taking an expectation over the trajectory information and only requiring conservativeness in expectation.

In this paper, we distinguish between the *observed prediction error* and *true prediction error*:

$$\mathbf{e}_i^{\text{obs}} := f_M\big(\mathbf{x}_i^{\text{obs}}\big) - \mathbf{y}_i^{\text{obs}} = \big[f_M\big(\mathbf{x}_i\big) - \mathbf{y}_i\big] + \big[f_M\big(\mathbf{x}_i + \epsilon_i^x\big) - f_M\big(\mathbf{x}_i\big)\big] - \epsilon_i^y, \tag{10}$$

$$\mathbf{e}_i^{\text{true}} := f_M\big(\mathbf{x}_i^{\text{obs}}\big) - \mathbf{y}_i = \big[f_M\big(\mathbf{x}_i\big) - \mathbf{y}_i\big] + \big[f_M\big(\mathbf{x}_i + \epsilon_i^x\big) - f_M\big(\mathbf{x}_i\big)\big]. \tag{11}$$

In practice, only the former can be computed from the observed trajectories. Since ϵ_i^x and ϵ_i^y are generated sequentially by the same estimator, they can be strongly correlated. Moreover, a perturbation in the model input can cause the model prediction to shift in the same direction through the learned dynamics. This creates a cancellation effect between the latter two parts in Equation (10), which can substantially reduce the magnitude of $\mathbf{e}_i^{\text{obs}}$. In such cases, calibration based on the observed prediction error will fail to provide valid coverage for the true prediction error.

4.2 INFORMATION-GUIDED AUGMENTATION

In this subsection, we leverage the available information in the current setting to construct the *augmented prediction error* and show that, under certain conditions, prediction sets constructed for the *augmented prediction error* achieve finite-sample marginal coverage of the *true prediction error*.

First, we define the *augmented prediction error* as:

$$\mathbf{e}_i^{\text{aug}} := f_M\big(\mathbf{x}_i^{\text{obs}}\big) - \mathbf{y}_i^{\text{obs}} + \xi_i = \mathbf{e}_i^{\text{obs}} + \xi_i, \tag{12}$$

where ξ_i is an artificially introduced random vector independently drawn from a zero-mean distribution with covariance $\mathbf{\Sigma}_i^{\xi}$, such as a Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_i^{\xi})$, and $\mathbf{\Sigma}_i^{\xi}$ satisfies the following assumption:

Assumption 2. *For $i = 1, \ldots, N+1$, $\mathbf{\Sigma}_i^{\xi}$ is an auxiliary covariance matrix constructed via the covariance mapping $\Psi\big(\mathbf{x}_i^{\text{obs}}, \mathbf{\Sigma}_i^{x,\text{obs}}, U_i; f_M\big) = \mathbf{\Sigma}_i^{\xi}$, and satisfies*

$$\mathbb{E}\Big[\mathbf{\Sigma}_i^{\xi}\Big] \succeq \text{Cov}\big(f_M(\mathbf{x}_i + \epsilon_i^x) - f_M(\mathbf{x}_i)\big), \tag{13}$$

where U_i represent i.i.d. auxiliary randomness, independent of all tuples $(\mathbf{x}_i, \mathbf{y}_i, \epsilon_i^x, \epsilon_i^y, \mathbf{\Sigma}_i^{x,\text{obs}})$.

Based on Assumption 1, Equation (13) essentially requires us to find a covariance mapping Ψ that conservatively propagates the uncertainty in the model input to the model output using the currently available information. For example, under Assumption 1, the following construction can be used:

i) If f_M is affine, we can construct $\mathbf{\Sigma}_i^{\xi} = \mathbf{J}\mathbf{\Sigma}_i^{x,\text{obs}}\mathbf{J}^\top$, where $\mathbf{J} = \partial f_M(\mathbf{x}_i^{\text{obs}})/\partial \mathbf{x}_i^{\text{obs}}$.

ii) If f_M is globally L-Lipschitz in the ℓ_2-norm, we can construct $\mathbf{\Sigma}_i^{\xi} = L^2\,\text{tr}(\mathbf{\Sigma}_i^{x,\text{obs}})\boldsymbol{I}_{d_y}$.

iii) If the state estimator further satisfies Equation (7) and ϵ_i^x is conditionally Gaussian given $\mathcal{I}_i$, then, for any measurable function f_M, we can construct

$$\mathbf{\Sigma}_i^{\xi} = \frac{1}{K}\sum_{m=1}^{K} \mathbf{d}_i^{(m)}(\mathbf{d}_i^{(m)})^\top, \tag{14}$$

where $\mathbf{d}_i^{(m)} = f_M(\mathbf{x}_i^{\text{obs}} + \widetilde{\epsilon}_i^{x,(m)}) - f_M\big(\mathbf{x}_i^{\text{obs}}\big)$, and $\widetilde{\epsilon}_i^{x,(m)}$, $m = 1, \ldots, K$, are i.i.d. samples from $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_i^{x,\text{obs}})$. Here, K can be any fixed positive integer.

The satisfaction of Assumption 2 by these constructions is proved in Appendix A. In fact, uncertainty propagation has been extensively studied (Monchot et al., 2023; Petersen et al., 2024; Kofnov et al., 2025), allowing us to flexibly adopt a variety of other propagation methods. In the experiments, we simply adopt the construction in Equation (14) and Jacobian-based propagation as practical approximations, which work well in practice and yield the desired empirical results.

The above procedure is designed to compensate for the last part in Equation (11) when cancellation occurs between the latter two parts of Equation (10). We then obtain the following result:

Theorem 1. *Under Assumptions 1 and 2, for each $i = 1, \ldots, N+1$, if $f_M(\mathbf{x}_i) - \mathbf{y}_i$ and ϵ_i^y are uncorrelated, then*

$$\mathbb{E}\big[e_i^{\text{aug}}\big] = \mathbb{E}\big[e_i^{\text{true}}\big], \quad \text{Cov}(\boldsymbol{e}_i^{\text{aug}}) \succeq \text{Cov}(\boldsymbol{e}_i^{\text{true}}). \tag{15}$$

This theorem establishes covariance dominance of the augmented prediction error over the true prediction error, with the proof provided in Appendix B. Notably, $f_M(\mathbf{x}_i) - \mathbf{y}_i$ and ϵ_i^y represent errors arising from two distinct systems—the trajectory world model and the state estimator, respectively. It is natural in practice for them to be uncorrelated, which is a weaker condition than independence.

Furthermore, the property established in Theorem 1 implies a dominance relation between the corresponding prediction sets when the prediction errors follow distributions of certain forms.

Theorem 2. *When Theorem 1 holds, suppose that, for each $i = 1, \ldots, N+1$, $\mathbf{e}_i^{\text{true}}$ and $\mathbf{e}_i^{\text{aug}}$ have centered Gaussian marginal distributions, then for any norm s,*

$$\mathbb{P}\Big(\mathbf{e}_{N+1}^{\text{true}} \in \widehat{\mathcal{E}}_{1-\alpha}\Big) \geq 1 - \alpha, \tag{16}$$

where

$$\widehat{\mathcal{E}}_{1-\alpha} := \big\{\boldsymbol{e} : s(\boldsymbol{e}) \leq \widehat{q}_{1-\alpha}\big\}, \tag{17}$$

and $\widehat{q}_{1-\alpha}$ is the $1-\alpha$ conformal quantile of $\{s(\mathbf{e}_i^{\text{aug}})\}_{i=1}^N$.

The proof of Theorem 2 and an additional sufficient condition for the non-Gaussian case are provided in Appendix B. The distributional requirement above provides a simple and interpretable sufficient condition, which can be further relaxed, as discussed in Appendix B.

4.3 NOISE-INFORMED CONFORMAL PREDICTION

Based on the above analysis, we propose Noise-Informed Conformal Prediction (NICP), which constructs prediction sets for the future states of the robotic system at each time step along a trajectory. Specifically, for a given time step t, NICP proceeds as follows:

i) Compute the model prediction error $\mathbf{e}_i^{\mathrm{obs}}(t)$ corresponding to the t-th time step of each calibration trajectory $\tau_1, \ldots, \tau_N$. **ii)** For each $i = 1, \ldots, N$, use a selected covariance mapping Ψ to propagate the covariance estimate $\mathbf{\Sigma}_i^{x,\mathrm{obs}}(t)$ provided by the state estimator, yielding an auxiliary covariance matrix $\mathbf{\Sigma}_i^{\xi}(t)$. **iii)** For each calibration trajectory, draw a centered Gaussian sample $\xi_i(t)$ with covariance $\mathbf{\Sigma}_i^{\xi}(t)$, and add it to $\mathbf{e}_i^{\mathrm{obs}}(t)$ to obtain the augmented prediction error $\mathbf{e}_i^{\mathrm{aug}}(t)$. **iv)** Compute the nonconformity scores $\{s(\mathbf{e}_i^{\mathrm{aug}}(t))\}_{i=1}^N$ under the selected norm s and obtain the conformal quantile $\widehat{q}_{1-\alpha}(t)$ at the target miscoverage level α. Then, output the prediction set

$$\widehat{\mathcal{C}}_{1-\alpha}^{(t)}(\mathbf{x}_{N+1}^{\mathrm{obs}}(t)) = \left\{\boldsymbol{y} : s\big(f_M(\mathbf{x}_{N+1}^{\mathrm{obs}}(t)) - \boldsymbol{y}\big) \le \widehat{q}_{1-\alpha}(t)\right\} \tag{18}$$

for the unknown state $\mathbf{y}_{N+1}(t)$ of a new trajectory τ_{N+1}. When Theorem 2 holds, this set satisfies

$$\mathbb{P}\Big(\mathbf{y}_{N+1}(t) \in \widehat{\mathcal{C}}_{1-\alpha}^{(t)}(\mathbf{x}_{N+1}^{\mathrm{obs}}(t))\Big) \ge 1 - \alpha. \tag{19}$$

This framework is flexible: for high-dimensional multi-step outputs of trajectory world models, the outputs can be treated either jointly or separately across prediction steps and state dimensions by regarding the quantity of interest as the target $\mathbf{y}$. In addition, the assumptions made in our analysis may not hold exactly in practice. For example, the state estimator may not be conservative, and the chosen covariance mapping Ψ may not strictly satisfy Assumption 2. Nevertheless, Theorems 1 and 2 can still hold under certain deviations from these assumptions. In fact, as shown in Appendix B, Theorem 1 only requires Equation (47) to hold. Since the first term in Equation (47) is always positive semidefinite, it is not necessary to have $\mathrm{Cov}(f_M(\mathbf{x}_i) - \mathbf{y}_i, \epsilon_i^y) = \mathbf{0}$, nor does Equation (13) need to hold exactly. It is sufficient that their combined effect satisfies the condition in Equation (47). This relaxation provides a certain degree of robustness to violations of the assumptions, while any positive slack also introduces additional conservativeness.

5 EXPERIMENTS

In this section, we evaluate the performance of the proposed method across different trajectory world models on a range of simulated and real-world datasets. In the main experiments, we use Equation (14) with $K = 1$ as the covariance mapping Ψ and the ℓ_2-norm as the nonconformity score s. The detailed procedure is summarized in Algorithm 1 of Appendix C.

5.1 EXPERIMENTAL SETTINGS

We present the core experimental setup below. More details are provided in Appendix D.

Datasets i) Simulated humanoid robot dataset. We select 200 reference motions for each of the three tasks (*Crouch*, *Jog Arc*, and *Walk*) from the BONES-SEED humanoid dataset. We then use Sonic (Luo et al., 2026) to track these motions with a Unitree G1 humanoid in MuJoCo (Todorov et al., 2012) and obtain the corresponding full control-state trajectories. **ii)** Simulated quadruped robot dataset. This dataset also contains three tasks: *Walk Octagon*, *Walk Stairs*, and *Walk Straight*. For each task, we generate 200 rollouts of a Unitree Go1 robot in MuJoCo, following the simulation pipeline provided by Alvarez-Padilla et al. (2025). For both simulated datasets, we use OpenVINS (Geneva et al., 2020) to generate noisy observations of the base-related states, together with the corresponding covariance estimates. **iii)** Real-world quadruped robot dataset. We collect 120 trajectories from a Unitree Go2 robot using a waypoint-following controller. A contact-aided InEKF (Hartley et al., 2020) provides base-state estimates and their associated covariance estimates. A Vicon motion-capture system is used to obtain the ground-truth robot states.

Models We evaluate uncertainty quantification results on three different trajectory world models: TDM (Schubert et al., 2023), TrajWorld (Yin et al., 2025) and WestWorld (Wang et al., 2026). For all three models, we follow the training protocol described by Wang et al. (2026) to obtain the corresponding pretrained checkpoints and further fine-tune them on our training set. Specifically, we

Table 1: Experimental results on the simulated humanoid robot dataset across three trajectory world models (mean ± standard deviation). Red denotes empirical failure to achieve the target coverage, while **bold** marks the valid method with the smallest average set size.

Model	Method	Humanoid Crouch		Humanoid Jog Arc		Humanoid Walk	
		Coverage (Target 90%)	Set Size ↓	Coverage (Target 90%)	Set Size ↓	Coverage (Target 90%)	Set Size ↓
TDM	Vanilla CP	78.04 ± 0.65	0.744 ± 0.015	82.23 ± 0.82	0.826 ± 0.012	77.00 ± 0.86	0.751 ± 0.014
	Robust CP	98.41 ± 0.17	2.105 ± 0.026	98.53 ± 0.24	2.017 ± 0.060	98.36 ± 0.20	2.026 ± 0.043
	NICP (ours)	91.69 ± 0.61	$\mathbf{1.144 \pm 0.027}$	91.93 ± 0.61	$\mathbf{1.215 \pm 0.017}$	90.59 ± 0.58	$\mathbf{1.179 \pm 0.025}$
TrajWorld	Vanilla CP	66.62 ± 1.32	0.596 ± 0.013	68.74 ± 0.99	0.655 ± 0.011	63.29 ± 1.04	0.579 ± 0.014
	Robust CP	98.08 ± 0.25	2.207 ± 0.028	98.21 ± 0.28	2.102 ± 0.054	97.54 ± 0.33	2.091 ± 0.061
	NICP (ours)	91.94 ± 0.81	$\mathbf{1.214 \pm 0.032}$	92.07 ± 0.67	$\mathbf{1.273 \pm 0.030}$	90.13 ± 0.65	$\mathbf{1.219 \pm 0.042}$
WestWorld	Vanilla CP	69.64 ± 1.20	0.628 ± 0.014	71.44 ± 1.02	0.685 ± 0.010	68.86 ± 0.80	0.632 ± 0.016
	Robust CP	98.02 ± 0.25	2.249 ± 0.028	98.23 ± 0.27	2.068 ± 0.053	97.80 ± 0.30	2.079 ± 0.066
	NICP (ours)	91.88 ± 0.79	$\mathbf{1.198 \pm 0.032}$	91.92 ± 0.70	$\mathbf{1.255 \pm 0.027}$	90.13 ± 0.62	$\mathbf{1.207 \pm 0.037}$

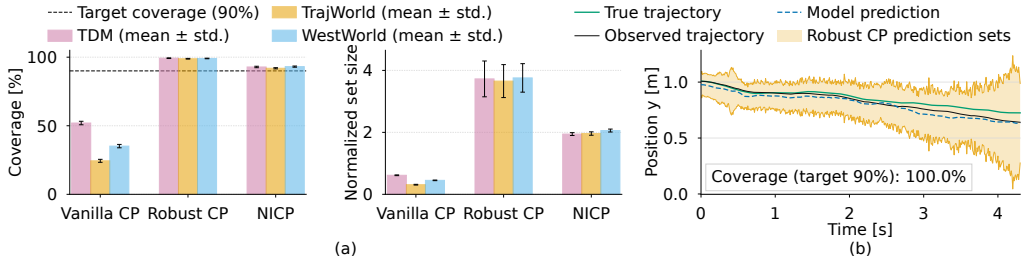


Figure 2: (a) Results on the real-world quadruped robot dataset. (b) Robust CP results for the y-axis position at each time step along the same trajectory as in Figure 1(b), with TDM used for prediction.

randomly select 50 trajectories per simulated task and 30 from the real-world dataset as the training sets for the corresponding tasks.

Baselines (1) Vanilla CP: Directly applies split conformal prediction (Lei et al., 2018) based on the observed prediction errors defined in Equation (10), representing the standard approach that ignores observation noise in trajectories. (2) Robust CP: A two-stage CP procedure is applied here, where a bounded CP set for the observation noise is first constructed using independently sampled noise based on the covariance estimates. An inflated threshold is then derived following the classical score-inflation paradigm introduced in Theorem 1 of Gendler et al. (2021), based on $s(\mathbf{e}_i^{\text{true}}) \leq s(\mathbf{e}_i^{\text{obs}}) + s(\epsilon_i^y)$. This represents an approach that handles the current scenario by treating the observation noise as bounded perturbations, a setting that has been widely studied.

Metrics We evaluate the empirical coverage and average set size for the system's true next-step states over all time steps of all test trajectories at $\alpha = 0.1$. Since different state dimensions represent different physical quantities, we construct prediction sets separately for each state dimension and average the results across them. Before averaging the set sizes across dimensions, we normalize the result for each dimension by that of Oracle CP, which directly uses the true prediction errors defined in Equation (11) to perform conformal prediction, thereby eliminating the effect of different physical units and scales. Additionally, since the state estimator only provides base-state estimates and their associated covariance estimates, our experiments focus only on these dimensions. We keep the other dimensions at their ground-truth values and set their error covariances to zero. Finally, the non-training trajectories are randomly split into calibration/test sets of 50/100 for each simulated dataset and 30/60 for the real-world dataset. All experiments are repeated over 10 random seeds.

5.2 Main Results

We evaluate the empirical coverage and average set size of the proposed method and the baselines across different trajectory world models and robotic tasks, as shown in Table 1 and Figure 2. The results show that **NICP achieves the target coverage of** 90%**, with empirical coverage closest to the target and the smallest average set size among the valid methods.** Similar results on the simulated quadruped robot dataset are reported in Table 7 of Appendix E. This suggests that,

Table 2: Experimental results for NICP using the Jacobian-based propagation as covariance mapping on the simulated humanoid robot dataset across three trajectory world models (mean ± std.).

Model	Method	Humanoid Crouch		Humanoid Jog Arc		Humanoid Walk	
		Coverage (Target 90%)	Set Size ↓	Coverage (Target 90%)	Set Size ↓	Coverage (Target 90%)	Set Size ↓
TDM	NICP-J	93.60 ± 0.50	1.322 ± 0.026	94.30 ± 0.47	1.422 ± 0.020	93.63 ± 0.46	1.439 ± 0.034
TrajWorld	NICP-J	93.22 ± 0.69	1.385 ± 0.036	93.26 ± 0.59	1.358 ± 0.039	92.17 ± 0.55	1.389 ± 0.053
WestWorld	NICP-J	92.98 ± 0.69	1.286 ± 0.030	92.71 ± 0.66	1.295 ± 0.031	92.29 ± 0.56	1.333 ± 0.045

Table 3: Experimental results across different prediction steps on the Humanoid Crouch task using the TDM model (mean ± standard deviation). Red denotes empirical failure to achieve the target coverage, while **bold** marks the valid method with the smallest average set size.

Method	Prediction Step 2		Prediction Step 3		Prediction Step 4	
	Coverage (Target 90%)	Set Size ↓	Coverage (Target 90%)	Set Size ↓	Coverage (Target 90%)	Set Size ↓
Vanilla CP	81.72 ± 0.71	0.815 ± 0.014	83.76 ± 0.76	0.851 ± 0.013	85.20 ± 0.77	0.875 ± 0.013
Robust CP	98.26 ± 0.21	2.005 ± 0.025	98.15 ± 0.23	1.942 ± 0.025	98.06 ± 0.25	1.892 ± 0.022
NICP (ours)	91.97 ± 0.63	$\mathbf{1.132 \pm 0.023}$	92.00 ± 0.68	$\mathbf{1.122 \pm 0.022}$	92.03 ± 0.70	$\mathbf{1.116 \pm 0.021}$

when applied to downstream tasks, NICP can provide prediction sets for the robotic system's true future states that achieve the desired uncertainty level without introducing excessive conservatism. In contrast, Vanilla CP fails to construct valid prediction sets that meet the target coverage, whereas Robust CP exhibits substantial overcoverage, resulting in excessively large set sizes. This is because NICP is developed based on the problem formulation and theoretical analysis in Section 4, which aligns well with our task scenario, while *the state estimators used in our experiments are reasonably consistent or conservative in aggregate*, as supported by the results in Table 6 and Appendix E. By contrast, Vanilla CP essentially calibrates the observed future states using the observed prediction error, while Robust CP requires the noise to be bounded before it can be applied, introducing additional conservatism in this scenario. Figure 1(b) and Figure 2(b) visualize the prediction sets produced by different CP methods, allowing for an intuitive comparison of their differences. Additional visualization results across other state dimensions are provided in Figure 10 of Appendix E.

5.3 Ablation Study

We also study the effects of different choices of the covariance mapping, prediction horizon, and state-estimator consistency level on the experimental results.

Effect of covariance mapping. As discussed in Section 4, the covariance mapping Ψ used in NICP can be flexibly chosen in practical applications. In addition to the construction used in the main experiments, another practical choice is $\Psi = \mathbf{J}_i \boldsymbol{\Sigma}_i^{x,\mathrm{obs}} \mathbf{J}_i^\top$, where $\mathbf{J}_i = \partial f_M(\mathbf{x}_i^{\mathrm{obs}})/\partial \mathbf{x}_i^{\mathrm{obs}}$. We refer to this variant as NICP-J, with the detailed procedure provided in Algorithm 2 of Appendix C. We evaluate NICP-J on the simulated humanoid robot dataset across different trajectory world models, with the results reported in Table 2. As shown, NICP-J is also valid but yields more conservative coverage than NICP in Table 1. This is because this covariance mapping is more conservative for these tasks, as verified in Appendix E, where we also examine the effect of K in Equation (14).

Effect of prediction horizon. Trajectory world models can autoregressively perform multi-step prediction. We can similarly treat the true system states at the prediction step of interest as the target variable $\mathbf{y}$ and then follow the same CP procedure. Accordingly, for the TDM model on the Humanoid Crouch task, we separately construct prediction sets for the system states at prediction steps 2 – 4, with the results reported in Table 3. We observe similar behavior to the one-step results in Table 1. This is because both the effect of observation noise and the CP procedure remain essentially the same across different prediction steps. We can also construct a multi-step joint prediction set such that the true system states at different prediction steps are jointly covered by the corresponding prediction tube. This requires treating the prediction errors across different prediction steps as a single vector. To obtain a more appropriate prediction-set shape, we use a weighted ℓ_∞-norm as the nonconformity score, with the corresponding details provided in Appendix E. We evaluate the

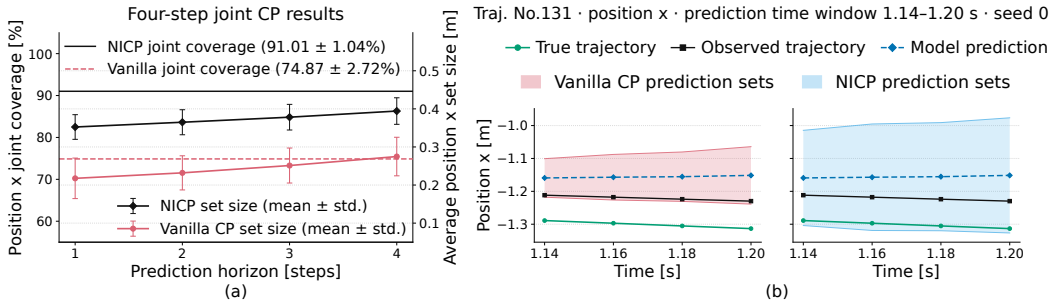


Figure 3: (a) Four-step joint conformal prediction results on the Humanoid Crouch task using the TDM model. (b) CP prediction sets over a local prediction window along a specific trajectory.

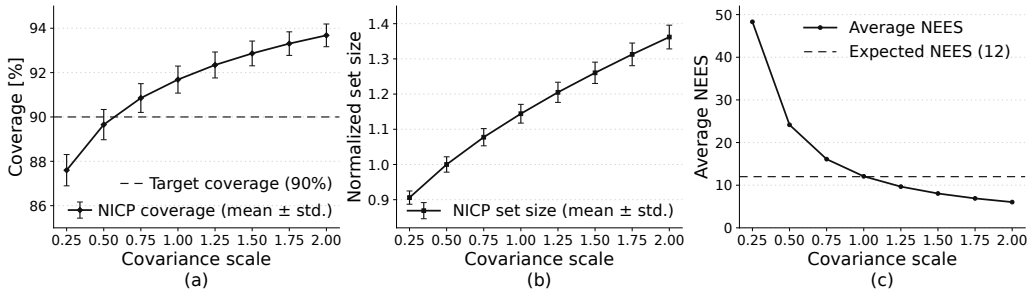


Figure 4: NICP results on the Humanoid Crouch task under different covariance scaling factors, with TDM used for prediction. (a) Empirical coverage. (b) Normalized prediction set size. (c) Average NEES computed using the scaled covariance estimates on the Humanoid Crouch task.

four-step joint CP performance for the x-axis position on the Humanoid Crouch task using the TDM model, with the results reported in Figure 3(a). The results show that NICP achieves the target joint coverage, while the prediction sets gradually become larger as the prediction horizon increases. This behavior is expected, as the model prediction error also increases with the prediction horizon. In contrast, Vanilla CP essentially calibrates to the observed future states and therefore fails to achieve the target coverage for the true future states. Figure 3(b) illustrates this failure mode through a specific example. Robust CP does not readily yield prediction sets with the same interpretation in this setting, so we do not extend it here.

Effect of estimator consistency level. The deployed state estimator may be miscalibrated and report relatively large or small covariance estimates. To investigate the impact of this on NICP, we uniformly scale the estimator-reported covariances on the Humanoid Crouch task. Figure 4 reports the CP results of the TDM model together with the corresponding NEES values, a standard estimator-consistency metric, computed using Equation (65). We observe that both the empirical coverage and set size of NICP increase monotonically with the covariance scaling factor. Conservative covariance estimates do not compromise the validity of NICP but lead to relatively higher coverage and larger prediction sets, while overly small covariance estimates can cause NICP to fail to achieve the target coverage. At the same time, a violation of estimator conservativeness does not immediately cause NICP to lose the target coverage. This is because the construction of the augmented prediction error introduces a certain degree of conservativeness, as discussed at the end of Section 4.

6 Conclusion and Limitations

In this paper, we proposed Noise-Informed Conformal Prediction (NICP), which can remain valid when all observations are corrupted by unbounded and correlated noise in the robotics pipeline. We also theoretically established its validity under certain conditions. Extensive experiments showed that, with a reasonably consistent state estimator, NICP can provide calibrated estimates of model uncertainty. One limitation of our method is that its theoretical guarantee relies on the quality of the covariance information provided by the state estimator and the form of the error distribution, which may be difficult to verify in practice.

AI USE STATEMENT

In this work, we used generative AI tools for polishing the writing, searching for relevant literature, providing coding assistance during method implementation, and assisting in the writing of proofs, specifically by checking mathematical derivations and improving their presentation. We did not use generative AI tools to generate synthetic datasets, clean or reformat datasets, develop theoretical models or conceptual frameworks, propose or refine hypotheses, provide critical ingredients for proving mathematical claims, formulate mathematical claims, design or provide feedback on research methodology or experiments, or interpret results. Qualitative and thematic data analysis is not applicable to this work. The research ideas, methodological and experimental designs, mathematical claims, and proof strategies were developed by the authors. All relevant literature was independently reviewed and verified by the authors and appropriately cited. All AI-assisted implementation code was checked by the authors. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide comprehensive dataset descriptions, implementation details, and evaluation protocols in Appendix D. Complete derivations and proofs of the theoretical results are included in Appendix A and Appendix B. The source code and datasets will be made publicly available upon acceptance.

REFERENCES

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021.

Mehran Aghabozorgi, Alireza Moazeni, Yanshu Zhang, and Ke Li. WIMLE: Uncertainty-aware world models with IMLE for sample-efficient continuous control. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=mzLOnTb3WH.

Juan Alvarez-Padilla, John Z Zhang, Sofia Kwok, John M Dolan, and Zachary Manchester. Real-time whole-body control of legged robots with model-predictive path integral control. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14721–14727. IEEE, 2025.

Theodore W Anderson. The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. *Proceedings of the American Mathematical Society*, 6(2): 170–176, 1955.

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

Yaakov Bar-Shalom, X Rong Li, and Thiagalingam Kirubarajan. *Estimation with applications to tracking and navigation: theory algorithms and software*. John Wiley & Sons, 2001.

Kong Yao Chee, M Ani Hsieh, and George J Pappas. Uncertainty quantification for learning-based mpc using weighted conformal prediction. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pp. 342–349. IEEE, 2023.

Chuchu Chen, Patrick Geneva, Yuxiang Peng, Woosik Lee, and Guoquan Huang. Optimization-based vins: Consistency, marginalization, and fej. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1517–1524. IEEE, 2023.

Chuchu Chen, Yuxiang Peng, and Guoquan Huang. Fast and consistent covariance recovery for sliding-window optimization-based vins. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13724–13731. IEEE, 2024.

Yahav Cohen, Jacob Goldberger, and Tom Tirer. Efficient conformal prediction for regression models under label noise. *arXiv preprint arXiv:2509.15120*, 2025.

Jacopo Diquigiovanni, Matteo Fontana, and Simone Vantini. Conformal prediction bands for multivariate functional data. *Journal of Multivariate Analysis*, 189:104879, 2022.

Anushri Dixit, Lars Lindemann, Skylar X Wei, Matthew Cleaveland, George J Pappas, and Joel W Burdick. Adaptive conformal prediction for motion planning among dynamic agents. In *Learning for Dynamics and Control Conference*, pp. 300–314. PMLR, 2023.

Bat-Sheva Einbinder, Shai Feldman, Stephen Bates, Anastasios N Angelopoulos, Asaf Gendler, and Yaniv Romano. Label noise robustness of conformal prediction. *Journal of Machine Learning Research*, 25(328):1–66, 2024.

Shai Feldman, Stephen Bates, and Yaniv Romano. Conformal prediction with corrupted labels: Uncertain imputation and robust re-weighting. In *International Conference on Learning Representations*, volume 2026, pp. 30901–30942, 2026.

Bernd Frauenknecht, Devdutt Subhasish, Friedrich Solowjow, and Sebastian Trimpe. On rollouts in model-based reinforcement learning. In *International Conference on Learning Representations*, volume 2025, pp. 76464–76489, 2025.

Asaf Gendler, Tsui-Wei Weng, Luca Daniel, and Yaniv Romano. Adversarially robust conformal prediction. In *International Conference on Learning Representations*, 2021.

Patrick Geneva, Kevin Eckenhoff, Woosik Lee, Yulin Yang, and Guoquan Huang. Openvins: A research platform for visual-inertial estimation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4666–4672. IEEE, 2020.

Subhankar Ghosh, Yuanjie Shi, Taha Belkhouja, Yan Yan, Jana Doppa, and Brian Jones. Probabilistically robust conformal prediction. In *Uncertainty in Artificial Intelligence*, pp. 681–690. PMLR, 2023.

Nicklas A Hansen, Hao Su, and Xiaolong Wang. Temporal difference learning for model predictive control. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8387–8406. PMLR, 17–23 Jul 2022.

Ross Hartley, Maani Ghaffari, Ryan M Eustice, and Jessy W Grizzle. Contact-aided invariant extended kalman filtering for robot state estimation. *The International Journal of Robotics Research*, 39(4):402–430, 2020.

Jianzhu Huai, Yukai Lin, Yuan Zhuang, and Min Shi. Consistent right-invariant fixed-lag smoother with application to visual inertial slam. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 6084–6092, 2021.

Linus Jeary, Tom Kuipers, Mehran Hosseini, and Nicola Paoletti. Verifiably robust conformal prediction. *Advances in Neural Information Processing Systems*, 37:4295–4314, 2024.

DR Jensen. Ordering ellipsoidal measures: Scale and peakedness orderings. *SIAM Journal on Applied Mathematics*, 44(6):1226–1231, 1984.

Andrey Kofnov, Daniel Kapla, Ezio Bartocci, and Efstathia Bura. Exact upper and lower bounds for the output distribution of neural networks with random inputs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 31133–31157. PMLR, 13–19 Jul 2025.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pp. 2796–2804. PMLR, 2018.

Abhijeet M Kulkarni, Ioannis Poulakakis, and Guoquan Huang. Proprioceptive-only state estimation for legged robots with set-coverage measurements of learned dynamics. *arXiv preprint arXiv:2603.18308*, 2026.

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*, 113 (523):1094–1111, 2018.

Chenhao Li, Andreas Krause, and Marco Hutter. Uncertainty-aware robotic world model makes offline model-based reinforcement learning work on real robots. *arXiv preprint arXiv:2504.16680*, 2025.

Daniil Lisus, Mitchell Cohen, and James Richard Forbes. Know what you don't know: Consistency in sliding window filtering with unobservable states applied to visual-inertial slam. *IEEE Robotics and Automation Letters*, 8(6):3382–3389, 2023.

Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Fernando Castañeda, Sirui Chen, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, et al. Sonic: Supersizing motion tracking for natural humanoid whole-body control. *Science Robotics*, 11(117):eaed4592, 2026.

Paul Monchot, Loic Coquelin, Sébastien Julien Petit, Sébastien Marmin, Erwan Le Pennec, and Nicolas Fischer. Input uncertainty propagation through trained neural networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25140–25173. PMLR, 23–29 Jul 2023.

Devesh Nath, Anutam Srinivasan, Haoran Yin, Ruitong Jiang, Jeffrey Fang, and Glen Chou. Pixels to proofs: Probabilistically-safe latent world model control via parallel conformal robust MPC. In *Robot World Models*, 2026. URL https://openreview.net/forum?id=PXF5zbGW3N.

Coby Penso and Jacob Goldberger. A conformal prediction score that is robust to label noise. *arXiv preprint arXiv:2405.02648*, 2024.

Felix Petersen, Aashwin Mishra, Hilde Kuehne, Christian Borgelt, Oliver Deussen, and Mikhail Yurochkin. Uncertainty quantification via stable distribution propagation. In *International Conference on Learning Representations*, volume 2024, pp. 10112–10138, 2024.

Ingmar Schubert, Jingwei Zhang, Jake Bruce, Sarah Bechtle, Emilio Parisotto, Martin Riedmiller, Jost Tobias Springenberg, Arunkumar Byravan, Leonard Hasenclever, and Nicolas Heess. A generalist dynamics model for control. *arXiv preprint arXiv:2305.10912*, 2023.

Junwon Seo, Kensuke Nakamura, and Andrea Bajcsy. Uncertainty-aware latent safety filters for avoiding out-of-distribution failures. In *9th Annual Conference on Robot Learning*, 2025. URL https://openreview.net/forum?id=CQKxhmLobo.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 5026–5033. IEEE, 2012.

Shenghua Wan, Le Gan, and De-Chuan Zhan. Learning to be uncertain: Pre-training world models with horizon-calibrated uncertainty. In *International Conference on Learning Representations*, volume 2026, pp. 106107–106127, 2026.

Yuchen Wang, Jiangtao Kong, Sizhe Wei, Xiaochang Li, Haohong Lin, Hongjue Zhao, Tianyi Zhou, Lu Gan, and Huajie Shao. Westworld: A knowledge-encoded scalable trajectory world model for diverse robotic systems. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=ncRRCG4BfP.

Skylar X Wei, Lu Gan, and Joel W Burdick. A safety-critical framework for ugvs in complex environments: A data-driven discrepancy-aware approach. *arXiv preprint arXiv:2403.03215*, 2024.

Shuo Yang, George J Pappas, Rahul Mangharam, and Lars Lindemann. Safe perception-based control under stochastic sensor uncertainty using conformal prediction. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pp. 6072–6078. IEEE, 2023.

Shaofeng Yin, Jialong Wu, Siqiao Huang, Xingjian Su, Xu He, Jianye HAO, and Mingsheng Long. Trajectory world models for heterogeneous environments. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Py2KmXaRmi.

A Covariance Mapping Construction

This appendix gives three covariance mappings Ψ that satisfy Assumption 2 under Assumption 1. For each $i = 1, \ldots, N+1$, define the model-output perturbation induced by the input observation noise as

$$\mathbf{u}_i := f_M\big(\mathbf{x}_i + \epsilon_i^x\big) - f_M(\mathbf{x}_i). \tag{20}$$

Assumption 1 gives

$$\mathbb{E}\big[\mathbf{\Sigma}_i^{x,\mathrm{obs}}\big] \succeq \mathrm{Cov}\big(\epsilon_i^x\big), \tag{21}$$

and we aim to construct a positive semidefinite random matrix $\mathbf{\Sigma}_i^{\xi}$ such that

$$\mathbb{E}\big[\mathbf{\Sigma}_i^{\xi}\big] \succeq \mathrm{Cov}(\mathbf{u}_i). \tag{22}$$

A.1 Linear Propagation

Proposition 1. *Suppose that the predictive model is affine,*

$$f_M(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \qquad \boldsymbol{A} \in \mathbb{R}^{d_y \times d_x}. \tag{23}$$

Then, for every $i = 1, \ldots, N+1$, the mapping

$$\mathbf{\Sigma}_i^{\xi} = \Psi\big(\mathbf{x}_i^{\mathrm{obs}}, \mathbf{\Sigma}_i^{x,\mathrm{obs}}, U_i; f_M\big) := \mathbf{J}\mathbf{\Sigma}_i^{x,\mathrm{obs}}\mathbf{J}^\top, \tag{24}$$

where $\mathbf{J} = \partial f_M(\mathbf{x}_i^{\mathrm{obs}})/\partial \mathbf{x}_i^{\mathrm{obs}} = \boldsymbol{A}$, satisfies Equation (22).

Proof. By Equation (23) and Equation (20),

$$\mathbf{u}_i = \boldsymbol{A}\epsilon_i^x, \tag{25}$$

and hence

$$\mathrm{Cov}(\mathbf{u}_i) = \boldsymbol{A}\mathrm{Cov}(\epsilon_i^x)\boldsymbol{A}^\top. \tag{26}$$

On the other hand, Equation (24) and Equation (21) give

$$\mathbb{E}[\mathbf{\Sigma}_i^{\xi}] = \boldsymbol{A}\mathbb{E}[\mathbf{\Sigma}_i^{x,\mathrm{obs}}]\boldsymbol{A}^\top \succeq \boldsymbol{A}\mathrm{Cov}(\epsilon_i^x)\boldsymbol{A}^\top. \tag{27}$$

Thus

$$\mathbb{E}\big[\mathbf{\Sigma}_i^{\xi}\big] \succeq \mathrm{Cov}(\mathbf{u}_i). \tag{28}$$

□

A.2 Lipschitz Construction

Proposition 2. *Suppose that $f_M : \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$ is globally L-Lipschitz with respect to the ℓ_2-norm. Then, for every $i = 1, \ldots, N+1$, the mapping*

$$\mathbf{\Sigma}_i^{\xi} = \Psi\big(\mathbf{x}_i^{\mathrm{obs}}, \mathbf{\Sigma}_i^{x,\mathrm{obs}}, U_i; f_M\big) := L^2\,\mathrm{tr}\big(\mathbf{\Sigma}_i^{x,\mathrm{obs}}\big)\boldsymbol{I}_{d_y} \tag{29}$$

satisfies Equation (22).

Proof. Fix i and let $\boldsymbol{a} \in \mathbb{R}^{d_y}$ be any unit vector. By the definition of covariance and the Cauchy–Schwarz inequality,

$$\boldsymbol{a}^\top\mathrm{Cov}(\mathbf{u}_i)\boldsymbol{a} = \mathrm{Var}(\boldsymbol{a}^\top\mathbf{u}_i) \le \mathbb{E}\big[(\boldsymbol{a}^\top\mathbf{u}_i)^2\big] \le \mathbb{E}\big[\|\mathbf{u}_i\|_2^2\big]. \tag{30}$$

The Lipschitz property and $\mathbb{E}[\epsilon_i^x] = \mathbf{0}$ yield

$$\mathbb{E}\big[\|\mathbf{u}_i\|_2^2\big] \le L^2\mathbb{E}\big[\|\epsilon_i^x\|_2^2\big] = L^2\,\mathrm{tr}\big(\mathrm{Cov}(\epsilon_i^x)\big) \le L^2\,\mathrm{tr}\big(\mathbb{E}[\mathbf{\Sigma}_i^{x,\mathrm{obs}}]\big), \tag{31}$$

where the last inequality follows from Equation (21).

By Equation (29) and $\|\boldsymbol{a}\|_2 = 1$, the last quantity in Equation (31) equals $\boldsymbol{a}^\top\mathbb{E}[\mathbf{\Sigma}_i^{\xi}]\boldsymbol{a}$. Combining Equation (30) and Equation (31) therefore gives

$$\boldsymbol{a}^\top\mathrm{Cov}(\mathbf{u}_i)\boldsymbol{a} \le \boldsymbol{a}^\top\mathbb{E}[\mathbf{\Sigma}_i^{\xi}]\boldsymbol{a}. \tag{32}$$

Since this holds for every unit vector $\boldsymbol{a}$, Equation (22) follows. □

A.3 Monte Carlo Construction

Proposition 3. *Suppose that the state estimator satisfies Equation (7) and that ϵ_i^x is conditionally Gaussian given the current information $\mathcal{I}_i$, which includes $\mathbf{x}_i^{\mathrm{obs}}$ and $\mathbf{\Sigma}_i^{x,\mathrm{obs}}$. For any measurable predictor f_M and any fixed integer $K \geq 1$, the covariance mapping*

$$\mathbf{\Sigma}_i^{\xi} = \Psi\big(\mathbf{x}_i^{\mathrm{obs}}, \mathbf{\Sigma}_i^{x,\mathrm{obs}}, U_i; f_M\big) := \frac{1}{K}\sum_{m=1}^{K} \mathbf{d}_i^{(m)} (\mathbf{d}_i^{(m)})^{\top} \tag{33}$$

satisfies Equation (22), where $\mathbf{d}_i^{(m)} = f_M(\mathbf{x}_i^{\mathrm{obs}} + \widetilde{\epsilon}_i^{x,(m)}) - f_M\big(\mathbf{x}_i^{\mathrm{obs}}\big)$ and

$$\widetilde{\epsilon}_i^{x,(m)} = (\mathbf{\Sigma}_i^{x,\mathrm{obs}})^{1/2} \boldsymbol{\eta}_i^{(m)}, \qquad m = 1, \ldots, K. \tag{34}$$

Here, $\boldsymbol{\eta}_i^{(m)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{d_x})$ are independent across i and m and independent of all trajectory data. The auxiliary randomness U_i in Equation (33) represents the collection $(\boldsymbol{\eta}_i^{(1)}, \ldots, \boldsymbol{\eta}_i^{(K)})$.

Proof. Conditional on the current information $\mathcal{I}_i$, both $\mathbf{x}_i^{\mathrm{obs}}$ and $\mathbf{\Sigma}_i^{x,\mathrm{obs}}$ are fixed. Therefore, under the above construction, $\widetilde{\epsilon}_i^{x,(m)}$ and ϵ_i^x have the same conditional distribution. Since $\mathbf{x}_i = \mathbf{x}_i^{\mathrm{obs}} - \epsilon_i^x$, we have

$$\mathbf{u}_i = f_M\big(\mathbf{x}_i^{\mathrm{obs}}\big) - f_M(\mathbf{x}_i^{\mathrm{obs}} - \epsilon_i^x), \tag{35}$$

which, together with the symmetry of the Gaussian distribution, implies

$$\mathbf{d}_i^{(m)} \mid \mathcal{I}_i \overset{d}{=} -\mathbf{u}_i \mid \mathcal{I}_i. \tag{36}$$

Consequently, for every fixed integer $K \geq 1$,

$$\mathbb{E}[\mathbf{\Sigma}_i^{\xi} \mid \mathcal{I}_i] = \frac{1}{K}\sum_{m=1}^{K} \mathbb{E}\big[\mathbf{d}_i^{(m)} (\mathbf{d}_i^{(m)})^{\top} \mid \mathcal{I}_i\big] = \mathbb{E}[\mathbf{u}_i \mathbf{u}_i^{\top} \mid \mathcal{I}_i]. \tag{37}$$

Taking expectations yields

$$\mathbb{E}[\mathbf{\Sigma}_i^{\xi}] = \mathbb{E}[\mathbf{u}_i \mathbf{u}_i^{\top}] = \mathrm{Cov}(\mathbf{u}_i) + \mathbb{E}[\mathbf{u}_i]\mathbb{E}[\mathbf{u}_i]^{\top} \succeq \mathrm{Cov}(\mathbf{u}_i). \tag{38}$$

Thus, the construction satisfies Equation (22) for every fixed $K \geq 1$. □

B Proof of Theorem 1 and Theorem 2

Let $\mathbf{z}_1, \ldots, \mathbf{z}_{N+1}$ be i.i.d. standard Gaussian random vectors in $\mathbb{R}^{d_y}$(or other random vectors in $\mathbb{R}^{d_y}$ with $\mathbb{E}[\mathbf{z}_i] = \mathbf{0}$ and $\mathrm{Cov}(\mathbf{z}_i) = \boldsymbol{I}_{d_y}$) that are independent of all trajectory tuples in Assumption 1 and all auxiliary randomness in Assumption 2. Define

$$\xi_i := (\mathbf{\Sigma}_i^{\xi})^{1/2} \mathbf{z}_i, \qquad i = 1, \ldots, N+1. \tag{39}$$

Then, conditional on the ith trajectory tuple and U_i, ξ_i has mean $\mathbf{0}$ and covariance $\mathbf{\Sigma}_i^{\xi}$. This construction also makes the error pairs $\{(\mathbf{e}_i^{\mathrm{true}}, \mathbf{e}_i^{\mathrm{aug}})\}_{i=1}^{N+1}$ i.i.d.

B.1 Proof of Theorem 1

Proof of Theorem 1. Fix i and define

$$\mathbf{m}_i := f_M(\mathbf{x}_i) - \mathbf{y}_i, \qquad \mathbf{u}_i := f_M(\mathbf{x}_i + \epsilon_i^x) - f_M(\mathbf{x}_i), \qquad \mathbf{v}_i := \epsilon_i^y. \tag{40}$$

Then

$$\mathbf{e}_i^{\mathrm{true}} = \mathbf{m}_i + \mathbf{u}_i, \qquad \mathbf{e}_i^{\mathrm{aug}} = \mathbf{m}_i + \mathbf{u}_i - \mathbf{v}_i + \xi_i. \tag{41}$$

Let $\mathcal{F}_i$ be the sigma-field generated by the ith trajectory tuple and the auxiliary randomness U_i. The covariance mapping in Assumption 2 makes $\mathbf{\Sigma}_i^{\xi}$ measurable with respect to $\mathcal{F}_i$.

From Equation (39),

$$\mathbb{E}[\xi_i \mid \mathcal{F}_i] = \mathbf{0}, \qquad \mathbb{E}[\xi_i \xi_i^{\top} \mid \mathcal{F}_i] = \mathbf{\Sigma}_i^{\xi}. \tag{42}$$

Consequently,

$$\mathbb{E}[\xi_i] = \mathbf{0}, \qquad \mathrm{Cov}(\xi_i) = \mathbb{E}[\mathbf{\Sigma}_i^{\xi}]. \tag{43}$$

Moreover, for every $\mathcal{F}_i$-measurable random vector $\mathbf{h}_i$,

$$\mathrm{Cov}(\mathbf{h}_i, \xi_i) = \mathbf{0}. \tag{44}$$

This is because $\mathbb{E}[\mathbf{h}_i \xi_i^\top] = \mathbb{E}[\mathbf{h}_i \mathbb{E}[\xi_i^\top \mid \mathcal{F}_i]] = \mathbf{0}$.

Using $\mathbb{E}[\mathbf{v}_i] = \mathbf{0}$ from Assumption 1, Equation (41) and Equation (43) give

$$\mathbb{E}[\mathbf{e}_i^{\mathrm{aug}}] - \mathbb{E}[\mathbf{e}_i^{\mathrm{true}}] = -\mathbb{E}[\mathbf{v}_i] + \mathbb{E}[\xi_i] = \mathbf{0}. \tag{45}$$

For the covariance, apply Equation (43) and Equation (44) with $\mathbf{h}_i = \mathbf{m}_i + \mathbf{u}_i - \mathbf{v}_i$ to obtain

$$\begin{aligned} &\mathrm{Cov}(\mathbf{e}_i^{\mathrm{aug}}) - \mathrm{Cov}(\mathbf{e}_i^{\mathrm{true}}) \\ &= \mathrm{Cov}(\mathbf{m}_i + \mathbf{u}_i - \mathbf{v}_i) - \mathrm{Cov}(\mathbf{m}_i + \mathbf{u}_i) + \mathbb{E}[\mathbf{\Sigma}_i^{\xi}] \\ &= \mathrm{Cov}(\mathbf{v}_i) - \mathrm{Cov}(\mathbf{m}_i + \mathbf{u}_i, \mathbf{v}_i) - \mathrm{Cov}(\mathbf{v}_i, \mathbf{m}_i + \mathbf{u}_i) + \mathbb{E}[\mathbf{\Sigma}_i^{\xi}]. \end{aligned} \tag{46}$$

By rearranging Equation (46), Equation (15) in Theorem 1 is equivalent to requiring

$$\mathrm{Cov}(\mathbf{u}_i - \mathbf{v}_i) - \Big(\mathrm{Cov}(\mathbf{m}_i, \mathbf{v}_i) + \mathrm{Cov}(\mathbf{v}_i, \mathbf{m}_i)\Big) + \Big(\mathbb{E}[\mathbf{\Sigma}_i^{\xi}] - \mathrm{Cov}(\mathbf{u}_i)\Big) \succeq \mathbf{0}. \tag{47}$$

In the above expression, the first term is a covariance matrix and is therefore positive semidefinite. The middle term vanishes by the uncorrelatedness condition in Theorem 1, and the last term is positive semidefinite by Assumption 2. This proves Theorem 1. □

B.2 Proof of Theorem 2

Lemma 1. *Let $s : \mathbb{R}^{d_y} \to [0, \infty)$ be a nonconformity score and let α be a given miscoverage level. Define*

$$\mathrm{r}_i^{\mathrm{true}} := s(\mathbf{e}_i^{\mathrm{true}}), \qquad \mathrm{r}_i^{\mathrm{aug}} := s(\mathbf{e}_i^{\mathrm{aug}}), \qquad i = 1, \ldots, N+1. \tag{48}$$

Suppose that

$$\mathbb{P}\big(\mathrm{r}_{N+1}^{\mathrm{true}} \le t\big) \ge \mathbb{P}\big(\mathrm{r}_{N+1}^{\mathrm{aug}} \le t\big), \qquad t \ge 0. \tag{49}$$

Let $k := \lceil (N+1)(1-\alpha) \rceil$ and let $\widehat{q}_{1-\alpha}$ be the kth order statistic of $\{\mathrm{r}_i^{\mathrm{aug}}\}_{i=1}^N \cup \{+\infty\}$. Then

$$\mathbb{P}\big(\mathbf{e}_{N+1}^{\mathrm{true}} \in \widehat{\mathcal{E}}_{1-\alpha}\big) \ge 1 - \alpha, \tag{50}$$

where

$$\widehat{\mathcal{E}}_{1-\alpha} := \{\boldsymbol{e} : s(\boldsymbol{e}) \le \widehat{q}_{1-\alpha}\}. \tag{51}$$

Proof. Define

$$F_{\mathrm{true}}(t) := \mathbb{P}(\mathrm{r}_{N+1}^{\mathrm{true}} \le t), \quad F_{\mathrm{aug}}(t) := \mathbb{P}(\mathrm{r}_{N+1}^{\mathrm{aug}} \le t), \tag{52}$$

where $F_{\mathrm{true}}(+\infty) = F_{\mathrm{aug}}(+\infty) = 1$. By Assumptions 1 and 2 and the construction in Equation (39), the error pairs, and hence the corresponding score pairs $\{(\mathrm{r}_i^{\mathrm{true}}, \mathrm{r}_i^{\mathrm{aug}})\}_{i=1}^{N+1}$, are i.i.d. Consequently, since the conformal quantile $\widehat{q}_{1-\alpha}$ is determined by the first N calibration scores, it is independent of $\mathrm{r}_{N+1}^{\mathrm{true}}$ and $\mathrm{r}_{N+1}^{\mathrm{aug}}$. Combining this independence with Equation (49), we obtain

$$\begin{aligned} \mathbb{P}\big(\mathrm{r}_{N+1}^{\mathrm{true}} \le \widehat{q}_{1-\alpha}\big) &= \mathbb{E}\big[F_{\mathrm{true}}(\widehat{q}_{1-\alpha})\big] \\ &\ge \mathbb{E}\big[F_{\mathrm{aug}}(\widehat{q}_{1-\alpha})\big] \\ &= \mathbb{P}\big(\mathrm{r}_{N+1}^{\mathrm{aug}} \le \widehat{q}_{1-\alpha}\big) \\ &\ge 1 - \alpha. \end{aligned} \tag{53}$$

By the definition of $\widehat{\mathcal{E}}_{1-\alpha}$, this proves the result. □

This lemma is adapted from Theorem 1 in Einbinder et al. (2024) to our specific setting, with Equation (49) being the key condition. In fact, at the target miscoverage level α, this condition can be relaxed to requiring

$$\mathbb{E}\big[F_{\mathrm{true}}(\widehat{q}_{1-\alpha})\big] \ge \mathbb{E}\big[F_{\mathrm{aug}}(\widehat{q}_{1-\alpha})\big]. \tag{54}$$

Based on Theorem 1, we next show that Equation (49), and hence the weaker condition in Equation (54), holds under certain error distributions, which leads to the proof of Theorem 2.

Proof of Theorem 2. Based on Theorem 1,

$$\mathrm{Cov}(\mathbf{e}_{N+1}^{\mathrm{aug}}) \succeq \mathrm{Cov}(\mathbf{e}_{N+1}^{\mathrm{true}}). \tag{55}$$

For $t \geq 0$, let

$$\mathcal{B}_t := \{\boldsymbol{z} \in \mathbb{R}^{d_y} : s(\boldsymbol{z}) \leq t\}. \tag{56}$$

Because s is a norm, $\mathcal{B}_t$ is centrally symmetric and convex. Additionally, the theorem assumes that the marginal distributions of $\mathbf{e}_{N+1}^{\mathrm{true}}$ and $\mathbf{e}_{N+1}^{\mathrm{aug}}$ are centered Gaussian. Therefore, applying Anderson's inequality (Anderson, 1955) together with Equation (55) yields

$$\mathbb{P}\big(\mathbf{e}_{N+1}^{\mathrm{true}} \in \mathcal{B}_t\big) \geq \mathbb{P}\big(\mathbf{e}_{N+1}^{\mathrm{aug}} \in \mathcal{B}_t\big), \qquad t \geq 0. \tag{57}$$

Equivalently,

$$\mathbb{P}\big(s(\mathbf{e}_{N+1}^{\mathrm{true}}) \leq t\big) \geq \mathbb{P}\big(s(\mathbf{e}_{N+1}^{\mathrm{aug}}) \leq t\big), \qquad t \geq 0. \tag{58}$$

Thus, Equation (49) holds. By Lemma 1, Theorem 2 follows. □

The same result applies beyond Gaussian errors when the two marginal distributions belong to a common centered elliptical family.

Corollary 1. *When Theorem 1 holds, suppose that, for each $i = 1, \ldots, N+1$, the marginal distributions of $\mathbf{e}_i^{\mathrm{true}}$ and $\mathbf{e}_i^{\mathrm{aug}}$ belong to a common centered elliptical family, which means*

$$\mathbf{e}_i^{\mathrm{true}} \overset{d}{=} \mathrm{r}\boldsymbol{M}_{\mathrm{true}}^{1/2}\mathbf{u}, \qquad \mathbf{e}_i^{\mathrm{aug}} \overset{d}{=} \mathrm{r}\boldsymbol{M}_{\mathrm{aug}}^{1/2}\mathbf{u}, \tag{59}$$

where $\mathrm{r} \geq 0$, $\mathbf{u} \sim \mathrm{Unif}(\mathbb{S}^{d_y-1})$, r and $\mathbf{u}$ are independent, and $\boldsymbol{M}_{\mathrm{true}}, \boldsymbol{M}_{\mathrm{aug}} \succeq \mathbf{0}$.

Then, for any norm s,

$$\mathbb{P}\big(\mathbf{e}_{N+1}^{\mathrm{true}} \in \widehat{\mathcal{E}}_{1-\alpha}\big) \geq 1 - \alpha, \tag{60}$$

where $\widehat{\mathcal{E}}_{1-\alpha}$ is defined in Equation (51).

Proof. Based on Theorem 1,

$$\mathrm{Cov}(\mathbf{e}_i^{\mathrm{aug}}) \succeq \mathrm{Cov}(\mathbf{e}_i^{\mathrm{true}}). \tag{61}$$

Since $\mathbb{E}[\mathbf{u}\mathbf{u}^\top] = \boldsymbol{I}_{d_y}/d_y$, Equation (59) implies

$$\mathrm{Cov}(\mathbf{e}_i^{\mathrm{true}}) = c_{\mathrm{r}}\boldsymbol{M}_{\mathrm{true}}, \qquad \mathrm{Cov}(\mathbf{e}_i^{\mathrm{aug}}) = c_{\mathrm{r}}\boldsymbol{M}_{\mathrm{aug}}, \qquad c_{\mathrm{r}} := \mathbb{E}[\mathrm{r}^2]/d_y. \tag{62}$$

If $c_{\mathrm{r}} = 0$, both distributions are degenerate at the origin and the result is immediate. Otherwise, Equation (61) gives

$$\boldsymbol{M}_{\mathrm{aug}} \succeq \boldsymbol{M}_{\mathrm{true}}. \tag{63}$$

The scale–peakedness ordering for centered ellipsoidal measures with a common radial generator (Jensen, 1984) then yields, for every centrally symmetric convex set $\mathcal{B} \subseteq \mathbb{R}^{d_y}$,

$$\mathbb{P}\big(\mathbf{e}_i^{\mathrm{true}} \in \mathcal{B}\big) \geq \mathbb{P}\big(\mathbf{e}_i^{\mathrm{aug}} \in \mathcal{B}\big). \tag{64}$$

Taking $\mathcal{B} = \mathcal{B}_t$ as defined in Equation (56), the above result implies Equation (49). Therefore, the condition of Lemma 1 is satisfied, and the desired conclusion follows. □

C ALGORITHMS FOR NOISE-INFORMED CONFORMAL PREDICTION

Algorithm 1 Noise-Informed Conformal Prediction (NICP)

Require: Pre-trained model f_M; time step t; norm s; miscoverage level α; calibration set $\mathcal{D}_{\text{cal}}(t) = \{(\mathbf{x}_i^{\text{obs}}(t), \mathbf{y}_i^{\text{obs}}(t), \mathbf{\Sigma}_i^{x,\text{obs}}(t))\}_{i=1}^N$;

for $i = 1, \ldots, N$ **do**
- $\widehat{\mathbf{y}}_i(t) \leftarrow f_M(\mathbf{x}_i^{\text{obs}}(t))$
- $\mathbf{e}_i^{\text{obs}}(t) \leftarrow \widehat{\mathbf{y}}_i(t) - \mathbf{y}_i^{\text{obs}}(t)$
- Sample $\widetilde{\epsilon}_i^x(t) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_i^{x,\text{obs}}(t))$
- $\mathbf{d}_i(t) \leftarrow f_M(\mathbf{x}_i^{\text{obs}}(t) + \widetilde{\epsilon}_i^x(t)) - \widehat{\mathbf{y}}_i(t)$
- $\mathbf{\Sigma}_i^{\xi}(t) \leftarrow \mathbf{d}_i(t)(\mathbf{d}_i(t))^\top$
- Sample $\xi_i(t) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_i^{\xi}(t))$
- $\mathbf{e}_i^{\text{aug}}(t) \leftarrow \mathbf{e}_i^{\text{obs}}(t) + \xi_i(t)$
- $\mathrm{r}_i^{\text{aug}}(t) \leftarrow s\big(\mathbf{e}_i^{\text{aug}}(t)\big)$

end for

$k \leftarrow \lceil (N+1)(1-\alpha) \rceil$

$\widehat{q}_{1-\alpha}(t) \leftarrow k$-th order statistic of $\{\mathrm{r}_i^{\text{aug}}(t)\}_{i=1}^N \cup \{+\infty\}$

For $\mathbf{x}_{N+1}^{\text{obs}}(t)$ at the new trajectory τ_{N+1}, **output**

$$\widehat{\mathcal{C}}_{1-\alpha}^{(t)}(\mathbf{x}_{N+1}^{\text{obs}}(t)) = \big\{\boldsymbol{y} : s\big(f_M(\mathbf{x}_{N+1}^{\text{obs}}(t)) - \boldsymbol{y}\big) \leq \widehat{q}_{1-\alpha}(t)\big\}.$$

Algorithm 2 Noise-Informed Conformal Prediction with Jacobian-based propagation (NICP-J)

Require: Pre-trained model f_M; time step t; norm s; miscoverage level α; calibration set $\mathcal{D}_{\text{cal}}(t) = \{(\mathbf{x}_i^{\text{obs}}(t), \mathbf{y}_i^{\text{obs}}(t), \mathbf{\Sigma}_i^{x,\text{obs}}(t))\}_{i=1}^N$;

for $i = 1, \ldots, N$ **do**
- $\widehat{\mathbf{y}}_i(t) \leftarrow f_M(\mathbf{x}_i^{\text{obs}}(t))$
- $\mathbf{e}_i^{\text{obs}}(t) \leftarrow \widehat{\mathbf{y}}_i(t) - \mathbf{y}_i^{\text{obs}}(t)$
- $\mathbf{J}_i(t) = \partial f_M(\mathbf{x}_i^{\text{obs}}(t)) / \partial \mathbf{x}_i^{\text{obs}}(t)$
- $\mathbf{\Sigma}_i^{\xi}(t) \leftarrow \mathbf{J}_i(t) \mathbf{\Sigma}_i^{x,\text{obs}}(t) \mathbf{J}_i(t)^\top$
- Sample $\boldsymbol{\xi}_i(t) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_i^{\xi}(t))$
- $\mathbf{e}_i^{\text{aug}}(t) \leftarrow \mathbf{e}_i^{\text{obs}}(t) + \boldsymbol{\xi}_i(t)$
- $\mathrm{r}_i^{\text{aug}}(t) \leftarrow s\big(\mathbf{e}_i^{\text{aug}}(t)\big)$

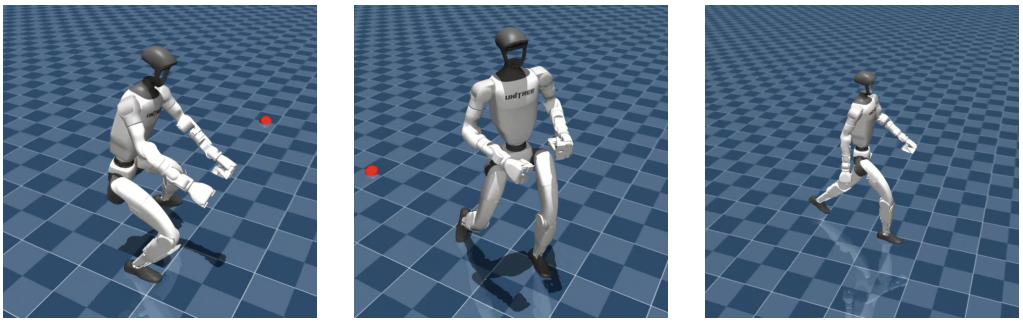
end for

$k \leftarrow \lceil (N+1)(1-\alpha) \rceil$

$\widehat{q}_{1-\alpha}(t) \leftarrow k$-th order statistic of $\{\mathrm{r}_i^{\text{aug}}(t)\}_{i=1}^N \cup \{+\infty\}$

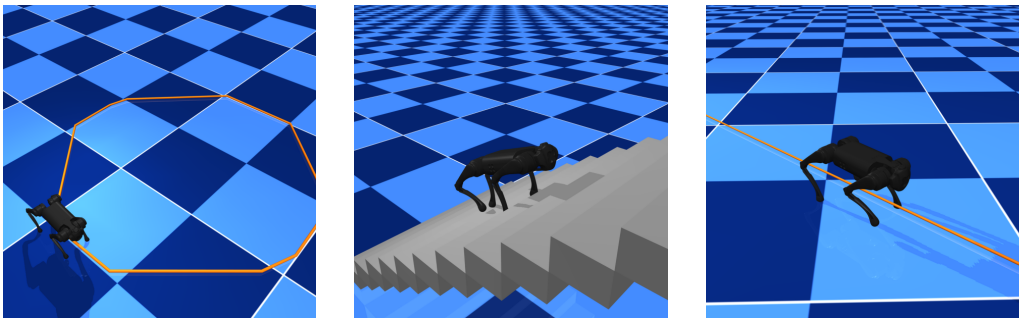
For $\mathbf{x}_{N+1}^{\text{obs}}(t)$ at the new trajectory τ_{N+1}, **output**

$$\widehat{\mathcal{C}}_{1-\alpha}^{(t)}(\mathbf{x}_{N+1}^{\text{obs}}(t)) = \big\{\boldsymbol{y} : s\big(f_M(\mathbf{x}_{N+1}^{\text{obs}}(t)) - \boldsymbol{y}\big) \leq \widehat{q}_{1-\alpha}(t)\big\}.$$



(a) Humanoid Crouch (b) Humanoid Jog Arc (c) Humanoid Walk

Figure 5: Illustrations of tasks in the simulated humanoid robot dataset.



(a) Quadruped Walk Octagon (b) Quadruped Walk Stairs (c) Quadruped Walk Straight

Figure 6: Illustrations of tasks in the simulated quadruped robot dataset.

D Detailed Experimental Settings

In this section, we provide a more detailed description of the experimental setup used in Section 5.

D.1 Data

i) Simulated humanoid robot dataset. BONES-SEED[1] is an open dataset of annotated human motion animations for humanoid robotics. We selected three tasks, *crouch*, *jog arc*, and *walk*, from the Locomotion package, as they exhibit substantial variation in the base-related states. For each task, we randomly selected 200 motion trajectories and tracked them at 50 Hz in MuJoCo (Todorov et al., 2012) using the Unitree G1 humanoid robot model and the Sonic whole-body control policy (Luo et al., 2026) with its default control pipeline. Figure 5 illustrates this process. At each time step, we recorded the joint control targets and robot states, yielding complete control-state trajectories.

ii) Simulated quadruped robot dataset. We generated the simulated quadruped trajectories in MuJoCo using the Unitree Go1 robot model and the whole-body MPPI locomotion controller provided by Alvarez-Padilla et al. (2025). We considered three waypoint-following tasks, *walk octagon*, *walk stairs*, and *walk straight*, and collected 200 rollouts per task at 100 Hz. Figure 6 illustrates this process. We retained the original MPPI controller and made only task-level modifications to obtain longer rollouts by extending the waypoint routes and, for *Walk Stairs*, the staircase accordingly. To introduce rollout diversity, we applied independent zero-mean Gaussian perturbations to the initial base position and generalized velocity. At each time step, we recorded the robot states and target joint positions, yielding complete control-state trajectories.

We use the simulator provided by OpenVINS (Geneva et al., 2020) to add observation noise to each simulated trajectory. For each trajectory, the simulator takes the complete *base position* and *base orientation* as input and outputs noisy *base position*, *base orientation*, *base linear velocity*, and *base*

[1] https://huggingface.co/datasets/bones-studio/seed

Table 4: OpenVINS simulation parameters used in our experiments. The upper section shows parameters shared by both simulated datasets, while the lower section lists dataset-specific settings for the simulated humanoid and quadruped robot datasets.

Parameter (shared)	**Value**	**Parameter (shared)**	**Value**
Accel. White Noise	6.0×10^{-3}	Gyro. White Noise	5.0904×10^{-4}
Accel. Random Walk	9.0×10^{-3}	Gyro. Random Walk	5.8179×10^{-5}
Camera Noise Std. (pixel)	1.5		
Parameter		**Humanoid**	**Quadruped**
Num. Clones		5	30
IMU Frequency (Hz)		50	100
Camera Frequency (Hz)		10	30
Feature Distance (m)		5–8	3–6

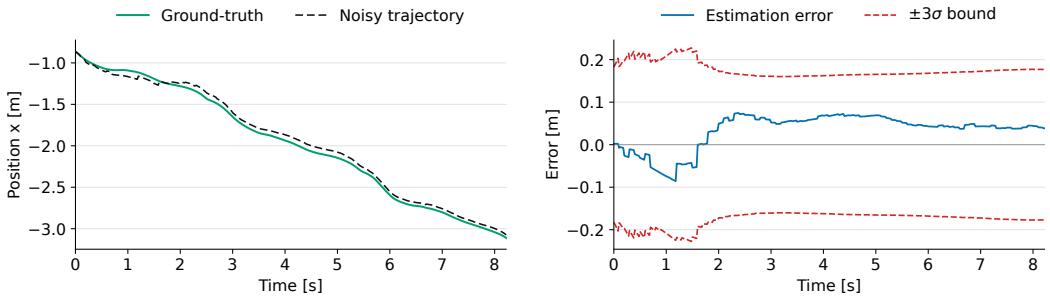


Figure 7: OpenVINS simulation results for trajectory No. 171 of the *Humanoid Crouch* task in the x-axis position. The left panel compares the ground-truth and noisy trajectories, while the right panel shows the simulated noise together with the 3σ error bounds derived from the covariance estimates provided by OpenVINS.

angular velocity, together with a corresponding 12×12 positive semidefinite covariance estimate of the observation noise. The *base orientation* is represented by a quaternion, while its covariance corresponds to a three-dimensional small-angle error representation. The simulator also outputs the corresponding ground-truth trajectory for these dimensions. This ground truth is reconstructed from the original trajectory via B-spline interpolation and is the trajectory actually used internally by the simulator; accordingly, the reported covariance estimates characterize the errors between this ground-truth trajectory and the noisy trajectory.

By default, the OpenVINS simulation pipeline only outputs data at the low-frequency camera timestamps. We therefore modify the output code to extract the internal estimates computed at the higher-frequency IMU timestamps, matching the sampling frequency of our data. For the angular-velocity covariance, we combine the gyroscope white-noise covariance with the filter-estimated gyroscope-bias covariance mapped through the corresponding Jacobian. We primarily follow the parameter configurations in the *rpng_sim* directory and *simulation.launch*, while adjusting several parameters, including the noise levels and IMU frequency, to suit our experimental setting. The modified parameters are listed in Table 4. Figure 7 illustrates the injected noise in the x-axis position along a specific task trajectory, together with the corresponding error estimates. Since the simulator only begins to simulate noise and produce outputs after the trajectory has accumulated sufficient displacement, the resulting trajectories are shorter than the original ones. For the simulated humanoid robot dataset, the average trajectory durations are $10.3\,\mathrm{s}$, $13.1\,\mathrm{s}$, and $7.5\,\mathrm{s}$ for *Crouch*, *Jog Arc*, and *Walk*, respectively. For the simulated quadruped robot dataset, the corresponding durations are $55.6\,\mathrm{s}$, $58.1\,\mathrm{s}$, and $55.5\,\mathrm{s}$ for *Walk Octagon*, *Walk Stairs*, and *Walk Straight*, respectively.

Finally, we align the original joint positions, joint velocities, and joint control targets with the newly generated noisy and ground-truth base trajectories, yielding complete paired noisy and ground-truth trajectories together with the corresponding covariance estimates of the observation noise. Table 5



Figure 8: Illustration of the real-world quadruped robot task.

Table 5: Data format used in our experiments, where J denotes the number of robot joints.

Type	**Dim**	**Component**	**Frame**	**Unit**
State	$[0{:}3]$	Base position	Global	m
	$[3{:}7]$	Base quaternion $(xyzw)$	Global to IMU	–
	$[7{:}10]$	Base linear velocity	IMU	m/s
	$[10{:}13]$	Base angular velocity	IMU	rad/s
	$[13{:}13+J]$	Joint position	–	rad
	$[13+J{:}13+2J]$	Joint velocity	–	rad/s
Action	$[0{:}J]$	Target joint position	–	rad
Covariance	12×12	Base-related covariance	–	–

summarizes the resulting data format, with $J = 29$ for the humanoid robot and $J = 12$ for the quadruped robot.

iii) Real-world quadruped robot dataset. We collected 120 real-world trajectories from a Unitree Go2 robot at approximately $100\,\mathrm{Hz}$ in a laboratory environment. For each trajectory, the robot followed an approximately $2.5\,\mathrm{m}$ waypoint route using a waypoint-following controller, resulting in trajectories with an average duration of $4.3\,\mathrm{s}$. Figure 8 illustrates this process. A contact-aided InEKF (Hartley et al., 2020) fused onboard IMU measurements with forward-kinematic corrections derived from joint-encoder measurements and detected foot contacts to estimate the base position, orientation, and linear velocity, together with their corresponding covariance estimates. The base angular velocity was obtained from gyroscope measurements corrected using the InEKF-estimated gyroscope bias, and its covariance was constructed by combining the estimated gyroscope-bias covariance with the gyroscope measurement-noise covariance. Meanwhile, a Vicon motion-capture system recorded the ground-truth base pose, with ground-truth velocities obtained by finite-differencing consecutive Vicon poses. At each time step, we also recorded the joint positions and velocities reported by the robot's low-level state interface and the target joint positions sent to the motors. The same joint-related states and actions are used in both the observed and ground-truth trajectories, yielding complete paired control-state trajectories. The resulting data follow the same format as in Table 5, with $J = 12$.

We also evaluate the consistency of the simulated and real-world state estimators using the Normalized Estimation Error Squared (NEES), a commonly used metric for assessing estimator consistency (Huai et al., 2021; Lisus et al., 2023; Chen et al., 2024). The NEES is computed over the base-related dimensions, with the orientation error converted from the quaternion representation to

Table 6: Consistency evaluation of the simulated and real-world state estimators on their respective datasets. For a consistent estimator, the expected average NEES is 12; lower values indicate conservativeness, while higher values indicate underestimated uncertainty.

Dataset	Task	Avg. NEES	Number of Points
Simulated Humanoid	Crouch	11.52	103,252
	Jog Arc	11.20	131,012
	Walk	11.57	75,681
Simulated Quadruped	Walk Octagon	11.57	1,111,592
	Walk Stairs	10.56	1,162,355
	Walk Straight	10.48	1,109,272
Real-world Quadruped	Walk	9.20	51,877

a three-dimensional small-angle error. Specifically, the NEES is computed as

$$\mathrm{NEES}_i(t) = \epsilon_i(t)^\top \left(\mathbf{\Sigma}_i^{\mathrm{obs}}(t)\right)^{-1} \epsilon_i(t), \tag{65}$$

where $\epsilon_i(t) \in \mathbb{R}^{12}$ denotes the observation noise at time step t of the i-th trajectory and $\mathbf{\Sigma}_i^{\mathrm{obs}}(t)$ is the corresponding covariance estimate. We average the NEES values over all time steps of all trajectories and report the results in Table 6. For a consistent state estimator, the average NEES is expected to be close to the error dimension, while a conservative estimator typically yields a smaller value; conversely, a larger NEES indicates underestimation of the observation uncertainty.

D.2 Models

We strictly followed the code and training settings provided by Wang et al. (2026) to obtain pretrained weights for the three trajectory world models. We then randomly selected 50 trajectories as the training set for each simulated task and 30 trajectories for the real-world task, with the remaining trajectories in each task used for calibration and testing. We initialized each model from its corresponding pretrained weights and fine-tuned it separately for each task on the corresponding training set for an additional 1,000 optimization steps using a fixed learning rate and a batch size of 16. The training trajectories were divided into non-overlapping temporal windows with a length and stride of 25, and prediction training was performed within each window. During inference on the calibration and test sets, the model used a one-step history to predict the future state at each time step along a trajectory. Both model training and inference are performed on the noisy trajectories. To accommodate our data structure, we set the maximum state and action dimensions supported by the models to 71 and 29, respectively. Additionally, for WestWorld, we constructed the required task-specific structural encodings from the MuJoCo model descriptions corresponding to the three robots in our experiments: the Unitree G1 29-DoF humanoid, the Unitree Go1 quadruped, and the Unitree Go2 quadruped. All other settings were kept consistent with the original code.

D.3 Baselines

For our baselines, Vanilla CP represents the standard approach that ignores observation noise in trajectories and calibrates directly on the observed model prediction errors defined in Equation (10). We directly apply split CP (Lei et al., 2018) to the nonconformity scores $\{s(\mathbf{e}_i^{\mathrm{obs}}(t))\}_{i=1}^N$ to obtain the corresponding conformal quantile and construct the prediction set accordingly.

Robust CP represents an approach that handles the current scenario by treating the observation noise as bounded perturbations. Existing robust CP methods (Gendler et al., 2021; Ghosh et al., 2023; Jeary et al., 2024) mainly consider settings where the model input or output is subject to bounded perturbations with a norm bound δ. Here, we follow the classical framework proposed by Gendler et al. (2021), whose key idea is to find a nonconformity score function $\tilde{s}$ satisfying:

$$\tilde{s}(\tilde{\mathbf{e}}) \leq \tilde{s}(\mathbf{e}) + m_\delta, \tag{66}$$

where $\tilde{\mathbf{e}}$ denotes the perturbed error, and $m_\delta \geq 0$ is a constant that is a function of δ. The robust prediction set can then be constructed as

$$\mathcal{C}^{\delta}_{1-\alpha} = \{\boldsymbol{y} : \tilde{s}(\tilde{\mathbf{e}}) \leq \tilde{q}_{1-\alpha} + m_\delta\}, \tag{67}$$

where $\tilde{q}_{1-\alpha}$ is the conformal quantile of $\{\tilde{s}(\mathbf{e})\}_{i=1}^{N}$.

In the setting considered in this paper, this construction can be carried out in a simple manner. Specifically, based on Equation (10) and Equation (11), the true prediction error can be expressed as

$$\mathbf{e}_i^{\text{true}} = \mathbf{e}_i^{\text{obs}} + \epsilon_i^y \tag{68}$$

Then, for any norm s, we have:

$$s(\mathbf{e}_i^{\text{true}}) \leq s(\mathbf{e}_i^{\text{obs}}) + s(\epsilon_i^y). \tag{69}$$

Therefore, to construct the set in Equation (67), we need to provide an upper bound for $s(\epsilon_i^y)$. This can be obtained through another conformal prediction procedure. Since ϵ_i^y represents the observation noise at time $t+1$, its covariance estimate is available on each calibration trajectory. We first sample the noise based on the corresponding covariance estimate for each calibration trajectory, and then use the conformal quantile of $\{s(\epsilon_i^y)\}_{i=1}^{N}$ as the value of m_δ in Equation (67). In this procedure, to achieve the final $1-\alpha$ coverage guarantee, we need to use a miscoverage level of $\alpha/2$ in each step, thereby introducing additional conservatism into the final prediction set.

We additionally perform Oracle CP to provide a normalization reference for the prediction-set sizes produced by the other methods. Since different output dimensions have different physical meanings and units, we need to normalize the average set size in each dimension before averaging across dimensions. Oracle CP directly uses ground-truth information to compute the conformal quantile of $\{s(\mathbf{e}_i^{\text{true}}(t))\}_{i=1}^{N}$ on the calibration trajectories and constructs prediction sets accordingly. It is worth noting that the true prediction error required by Oracle CP is unavailable in practical tasks.

D.4 METRICS

Trajectory world models can perform autoregressive multi-step predictions. Since the calibration trajectories are complete, we can compute multi-step prediction errors at each time step and calibrate them accordingly. In the main experiments, we only calibrate the one-step prediction results, while the calibration procedure for other prediction horizons is the same and is reported in the ablation study. In addition, different state dimensions represent different physical quantities, making a single joint prediction region across heterogeneous dimensions of limited practical meaning. Therefore, we construct prediction intervals independently for each output dimension. This only requires treating the prediction target $\mathbf{y}_{N+1}$ as the corresponding variable during conformal prediction and computing the associated quantities accordingly. Additionally, since the state estimator only provides base-state estimates and their associated covariance estimates, our experiments focus only on these dimensions. We keep the other dimensions at their ground-truth values and set their error covariances to zero.

For evaluation, the non-training trajectories are randomly split into calibration/test sets of 50/100 for each simulated task and 30/60 for the real-world task. For each base-related dimension, at a miscoverage level of $\alpha = 0.1$, we use the ℓ_2-norm as the nonconformity score and evaluate the empirical coverage and average prediction-set size over all time steps across all test trajectories. Tail time steps with insufficient calibration samples are excluded from the evaluation for all methods, accounting for 0.63% of all test prediction points. When constructing prediction sets for the base orientation, we convert the quaternion representation to roll-pitch-yaw angles, resulting in 12 base-related dimensions in total. We then average the results across these dimensions. Before averaging the set sizes, we normalize the average set size of each method in each dimension by the corresponding average set size of Oracle CP, such that a normalized set size of 1 corresponds to the Oracle CP set size. All experiments are repeated over 10 random seeds, and we report the mean and standard deviation across runs.

Table 7: Experimental results on the simulated quadruped robot dataset across three trajectory world models (mean ± standard deviation). Red denotes empirical failure to achieve the target coverage, while **bold** marks the valid method with the smallest average set size.

Model	Method	Quadruped Walk Octagon		Quadruped Walk Stairs		Quadruped Walk Straight	
		Coverage (Target 90%)	Set Size ↓	Coverage (Target 90%)	Set Size ↓	Coverage (Target 90%)	Set Size ↓
TDM	Vanilla CP	79.77 ± 0.28	0.828 ± 0.009	87.53 ± 0.31	0.952 ± 0.003	75.92 ± 0.38	0.774 ± 0.007
	Robust CP	98.73 ± 0.06	2.243 ± 0.056	97.87 ± 0.08	1.779 ± 0.020	98.92 ± 0.06	2.276 ± 0.030
	NICP (ours)	92.30 ± 0.15	$\mathbf{1.430 \pm 0.037}$	91.57 ± 0.18	$\mathbf{1.100 \pm 0.008}$	93.02 ± 0.18	$\mathbf{1.417 \pm 0.018}$
TrajWorld	Vanilla CP	64.49 ± 0.24	0.635 ± 0.006	77.04 ± 0.33	0.778 ± 0.006	52.56 ± 0.30	0.518 ± 0.004
	Robust CP	98.34 ± 0.12	2.486 ± 0.060	98.54 ± 0.04	2.329 ± 0.053	98.80 ± 0.12	2.215 ± 0.061
	NICP (ours)	92.98 ± 0.21	$\mathbf{1.590 \pm 0.042}$	93.64 ± 0.11	$\mathbf{1.545 \pm 0.036}$	93.24 ± 0.26	$\mathbf{1.518 \pm 0.045}$
WestWorld	Vanilla CP	69.11 ± 0.25	0.704 ± 0.008	80.26 ± 0.30	0.821 ± 0.005	57.15 ± 0.36	0.546 ± 0.005
	Robust CP	98.35 ± 0.11	2.471 ± 0.053	98.35 ± 0.05	2.393 ± 0.063	98.85 ± 0.11	2.265 ± 0.062
	NICP (ours)	92.22 ± 0.19	$\mathbf{1.533 \pm 0.038}$	93.14 ± 0.11	$\mathbf{1.462 \pm 0.032}$	93.05 ± 0.26	$\mathbf{1.510 \pm 0.044}$

Table 8: Dimension-wise average state-estimator NEES and empirical coverage for NICP on the Quadruped Walk Octagon task using the TDM model. For NEES, the reference value is 1, with lower values indicating conservative covariance estimates and higher values indicating that the state estimator underestimates uncertainty; the target coverage is 90%.

Axis	Position		Orientation		Linear Velocity		Angular Velocity	
	Avg. NEES	Coverage(%)	Avg. NEES	Coverage(%)	Avg. NEES	Coverage(%)	Avg. NEES	Coverage(%)
x / Roll	0.495	95.91 ± 0.45	0.671	92.35 ± 0.17	3.052	87.18 ± 0.34	1.011	90.53 ± 0.18
y / Pitch	1.378	91.66 ± 0.85	0.658	91.57 ± 0.24	0.887	91.06 ± 0.13	1.022	90.62 ± 0.19
z / Yaw	0.032	100.00 ± 0.00	0.080	95.65 ± 0.17	1.077	90.11 ± 0.25	0.910	90.93 ± 0.16

E Supplementary Experimental Results

In this section, we present the supplementary experimental results mentioned in Section 5.

E.1 Supplementary Main Experimental Results

Table 7 reports the experimental results on the simulated quadruped robot dataset across three trajectory world models. Similar to the results presented in Table 1 and Figure 2, NICP achieves the target coverage of 90% on average, with empirical coverage closest to the target and the smallest average set size among the valid methods. Table 8 provides a more fine-grained evaluation by reporting the dimension-wise empirical coverage of NICP on the Quadruped Walk Octagon task using the TDM model, together with the average NEES of the state-estimator covariance estimates computed separately for each dimension. Although the state estimator is reasonably consistent in aggregate, its estimation quality varies across individual dimensions. For example, the covariance estimate for the x-axis linear velocity is substantially underestimated, leading to undercoverage of NICP in this dimension. This observation is consistent with the effect of estimator consistency on NICP discussed in Section 5.3. Nevertheless, NICP remains valid and achieves empirical coverage close to the target level in most dimensions, demonstrating a considerable degree of robustness to variations in estimator consistency. In addition, since the state estimator is reasonably consistent or conservative in aggregate across all dimensions, the average coverage reported in Table 7 reaches the target level. Together with the results in Figure 4, this suggests that the effect of estimator consistency is similar at both the aggregate and individual-dimension levels.

Our method aims to establish valid coverage of the system's true future states under a given estimator behavior. Therefore, although the aggregate and dimension-wise results may differ to some extent, they exhibit a consistent pattern from the perspective of estimator consistency. In addition, Figure 10 provides visualization results for other state dimensions along the same real-world Unitree Go2 trajectory as in Figures 1(b) and 2(b). For clarity, we only show the CP results over the first 2 seconds for the x-axis position, roll orientation, x-axis linear velocity, and x-axis angular velocity. The results are consistent with the discussion in Section 5.

Table 9: Experimental results of NICP on the Humanoid Crouch task across three trajectory world models under different choices of K in Equation (14) (mean $\pm$ standard deviation).

Model	Method	$K = 4$		$K = 8$		$K = 16$	
		Coverage (Target 90%)	**Set Size ↓**	**Coverage (Target 90%)**	**Set Size ↓**	**Coverage (Target 90%)**	**Set Size ↓**
TDM	NICP	91.97 ± 0.60	1.158 ± 0.022	92.01 ± 0.61	1.161 ± 0.023	92.02 ± 0.59	1.161 ± 0.022
TrajWorld	NICP	92.19 ± 0.79	1.226 ± 0.027	92.22 ± 0.80	1.228 ± 0.029	92.21 ± 0.79	1.227 ± 0.028
WestWorld	NICP	92.17 ± 0.78	1.214 ± 0.026	92.19 ± 0.80	1.216 ± 0.027	92.20 ± 0.77	1.215 ± 0.025

Table 10: Conservativeness evaluation of the covariance mappings used by NICP and NICP-J across three Humanoid tasks and three trajectory world models based on Equation (70). Values below 12 indicate conservativeness in aggregate.

Model	**Humanoid Crouch**		**Humanoid Jog Arc**		**Humanoid Walk**	
	NICP	**NICP-J**	**NICP**	**NICP-J**	**NICP**	**NICP-J**
TDM	6.401	3.481	9.687	4.386	9.470	5.227
TrajWorld	6.420	4.271	9.614	6.585	9.701	6.512
WestWorld	6.643	2.573	9.922	6.695	9.663	5.332

E.2 Supplementary Validation Results

In Appendix A, we show that the choice of K in Equation (14) does not affect the validity of Proposition 3. However, this result relies on Equation (7) holding exactly. In practical settings, it is unclear how the choice of K affects the final conformal prediction results. To investigate this, we evaluate different choices of K on the Humanoid Crouch task across all three trajectory world models and report the results in Table 9. As shown, the empirical coverage and average prediction-set size vary only slightly over the tested values of K, indicating limited empirical sensitivity in this setting. One possible explanation is that, although K affects the stability of each individual auxiliary covariance matrix, the conformal quantile aggregates information across many calibration samples and is therefore less sensitive to fluctuations in individual samples.

The NICP results reported in Tables 1 and 2 demonstrate the practical effectiveness of the two corresponding covariance-mapping constructions. We further analyze the relevant steps in the algorithmic pipeline. The results in Table 6 indicate that the state estimator is reasonably consistent across the three Humanoid tasks. *Since the effect of estimator consistency on the final results exhibits similar behavior at both the aggregate and individual-dimension levels, we mainly conduct the following analysis on the full error vector to facilitate comparison with the estimator-consistency results and the main experimental results.* Specifically, we similarly compute the NEES as

$$\mathrm{NEES}_i(t) = \mathbf{u}_i(t)^\top \left(\overline{\mathbf{\Sigma}^{\xi}}\right)^{-1} \mathbf{u}_i(t), \tag{70}$$

where $\mathbf{u}_i(t) = f_M(\mathbf{x}_i^{\mathrm{obs}}(t)) - f_M(\mathbf{x}_i(t))$, and $\overline{\mathbf{\Sigma}^{\xi}}$ denotes the mean of $\mathbf{\Sigma}_i^{\xi}(t)$ over all trajectories τ_i and time steps t. We use the averaged auxiliary covariance matrix because the auxiliary covariance matrices may be singular. While this does not affect the sampling procedure, it prevents their direct use in the NEES calculation. Moreover, Assumption 2 only constrains the expectation of the auxiliary covariance matrix. Table 10 reports the NEES results of the two covariance mappings across three trajectory world models and three Humanoid tasks. We observe that both mappings exhibit a certain degree of conservativeness in aggregate, with Jacobian-based propagation being more conservative than the sampling-based approach. This explains why NICP-J yields more conservative coverage and larger prediction sets in Table 2 than NICP in Table 1.

We further compute, over all trajectories and time steps, the ℓ_2-norm of the mean error vector and the dimension-averaged variance ratios between each type of error and the true prediction error. The results are reported in Table 11. We observe that the mean vectors of all error types are very close to zero. In addition, in aggregate, the augmented prediction errors corresponding to both covariance

Table 11: The ℓ_2-norms of the mean errors and the dimension-averaged variance ratios relative to the true prediction error across three trajectory world models and Humanoid tasks.

Model	Error	Humanoid Crouch		Humanoid Jog Arc		Humanoid Walk	
		$\lvert\mathbb{E}[\mathbf{e}]\rvert_2$	Variance ratio	$\lvert\mathbb{E}[\mathbf{e}]\rvert_2$	Variance ratio	$\lvert\mathbb{E}[\mathbf{e}]\rvert_2$	Variance ratio
TDM	$\mathbf{e}^{\text{obs}}$	0.009	0.650	0.015	0.779	0.018	0.616
	$\mathbf{e}^{\text{true}}$	0.005	1.000	0.018	1.000	0.017	1.000
	$\mathbf{e}^{\text{aug}}$(NICP)	0.009	1.586	0.015	1.236	0.018	1.534
	$\mathbf{e}^{\text{aug}}$(NICP-J)	0.009	2.309	0.015	1.746	0.017	2.472
TrajWorld	$\mathbf{e}^{\text{obs}}$	0.008	0.512	0.007	0.611	0.012	0.447
	$\mathbf{e}^{\text{true}}$	0.008	1.000	0.009	1.000	0.017	1.000
	$\mathbf{e}^{\text{aug}}$(NICP)	0.008	1.816	0.007	1.313	0.012	1.674
	$\mathbf{e}^{\text{aug}}$(NICP-J)	0.008	2.407	0.007	1.601	0.012	2.201
WestWorld	$\mathbf{e}^{\text{obs}}$	0.005	0.663	0.012	0.631	0.013	0.552
	$\mathbf{e}^{\text{true}}$	0.007	1.000	0.014	1.000	0.020	1.000
	$\mathbf{e}^{\text{aug}}$(NICP)	0.005	1.658	0.012	1.273	0.013	1.438
	$\mathbf{e}^{\text{aug}}$(NICP-J)	0.005	3.902	0.012	1.476	0.013	2.675

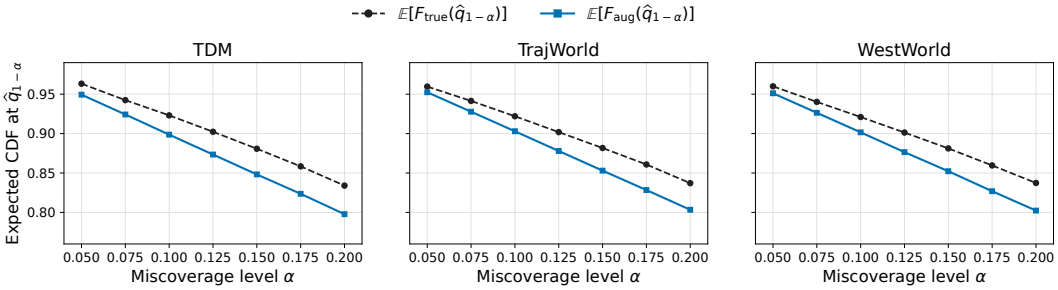


Figure 9: Empirical verification of Equation (54) under NICP augmentation on the Humanoid Crouch task across three trajectory world models.

mappings have larger variances than the true prediction errors, whereas the observed prediction errors have substantially smaller variances than the true prediction errors. These results support the effectiveness of our augmented-error construction and provide empirical evidence that observation noise induces a cancellation effect in the prediction error. The results in Figure 9 further empirically validate Equation (54). Specifically, across 10 random seeds, we examine the expected Cumulative Distribution Function (CDF) dominance of the error vectors under the ℓ_2-norm on the Humanoid Crouch task. Across the tested miscoverage levels α, the estimated expected CDF values of the true-error nonconformity scores at $\hat{q}_{1-\alpha}$ are larger than those of the augmented-error scores. These results provide empirical support for the sufficient condition under which prediction sets calibrated on the augmented prediction errors attain the target marginal coverage $1-\alpha$ for the true prediction errors.

E.3 Nonconformity Score for Multi-Step Prediction

In this subsection, we provide a detailed description of the nonconformity score used in Section 5.3 for constructing the multi-step joint prediction sets. We can simply treat the multi-step prediction errors as a single vector and use the ℓ_2-norm as the nonconformity score, which is theoretically valid. However, the resulting prediction-set shape may be undesirable, as it ignores the differences in error magnitude across prediction steps. Therefore, we instead use a specially designed norm as the nonconformity score to obtain a more appropriate prediction-set shape.

Specifically, for a given target state dimension, time step, and prediction horizon M, let $\mathrm{e}_i^{\text{obs},j}$, $j = 1,\ldots,M$, denote the observed prediction error at the j-th prediction step, and stack them as $\mathbf{e}_i^{\text{obs}} = [\mathrm{e}_i^{\text{obs},1},\ldots,\mathrm{e}_i^{\text{obs},M}]^\top$. Similarly, following the NICP procedure, we obtain the corresponding $\mathbf{e}_i^{\text{aug}} = [\mathrm{e}_i^{\text{aug},1},\ldots,\mathrm{e}_i^{\text{aug},M}]^\top$. Inspired by Diquigiovanni et al. (2022), we can compute the nonconformity

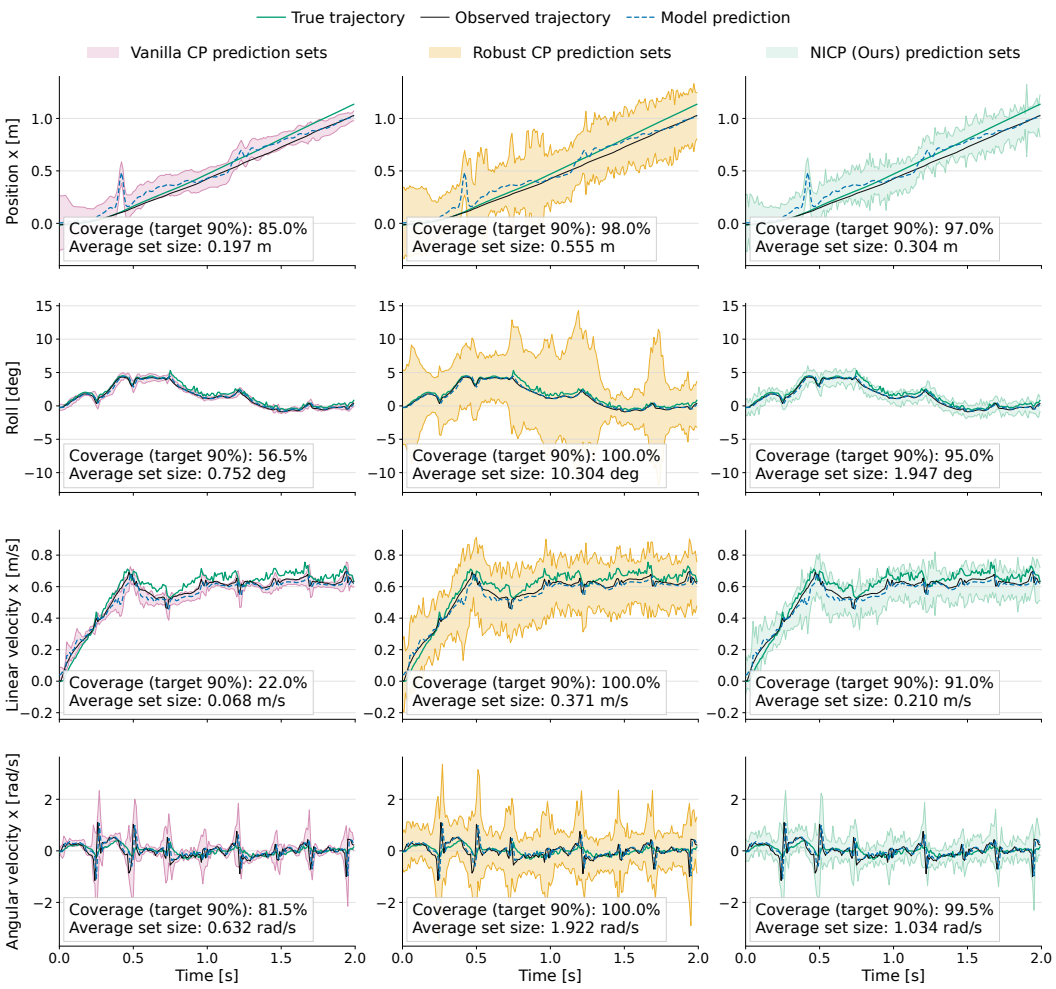


Figure 10: Conformal prediction results for four base-state dimensions over the first 2 seconds of the same real-world Unitree Go2 trajectory as in Figure 1(b), with TDM used for one-step prediction.

scores for Vanilla CP and NICP, respectively, as

$$s_1(\mathbf{e}_i^{\text{obs}}) = \max_{j=1,\dots,M} \frac{|\mathbf{e}_i^{\text{obs},j}|}{\widehat{\sigma}^{\text{obs},j}}, \quad s_2(\mathbf{e}_i^{\text{aug}}) = \max_{j=1,\dots,M} \frac{|\mathbf{e}_i^{\text{aug},j}|}{\sqrt{(\widehat{\sigma}^{\text{obs},j})^2 + \overline{\Sigma}^{\xi}(j,j)}}. \tag{71}$$

Here, both s_1 and s_2 can be shown to be norms on $\mathbb{R}^M$. The quantity $\widehat{\sigma}^{\text{obs},j}$ denotes the sample standard deviation of $\mathbf{e}_i^{\text{obs},j}$, while $\overline{\Sigma}^{\xi}(j,j)$ denotes the sample mean of the j-th diagonal element of the auxiliary covariance matrix $\mathbf{\Sigma}_i^{\xi}$; both are estimated using a held-out subset of the calibration set, thereby reducing the effective calibration-set size. For the results in Figure 3, we hold out 10 trajectories from the 50 calibration trajectories for parameter estimation and compute the nonconformity scores using the remaining 40 trajectories. In this case, the prediction-set sizes at different prediction steps can be flexibly adjusted according to their respective error scales without compromising the joint coverage. This approach can also be applied to construct joint prediction sets for other quantities of interest.