DoublyAware: Dual Planning and Policy Awareness for Temporal Difference Learning in Humanoid Locomotion

Khang Nguyen[1], An T. Le[2], Jan Peters[2,3,4], and Minh Nhat Vu[5,6]

***Abstract*—Achieving robust robot learning for humanoid locomotion is a fundamental challenge in model-based reinforcement learning (MBRL), where environmental stochasticity and randomness can hinder efficient exploration and learning stability. The environmental, so-called aleatoric, uncertainty can be amplified in high-dimensional action spaces with complex contact dynamics, and further entangled with epistemic uncertainty in the models during learning phases. In this work, we propose *DoublyAware*, an uncertainty-aware extension of Temporal Difference Model Predictive Control (TD-MPC) that explicitly decomposes uncertainty into two disjoint interpretable components, *i.e.*, planning and policy uncertainties. To handle the planning uncertainty, *DoublyAware* employs conformal prediction to filter candidate trajectories using quantile-calibrated risk bounds, ensuring statistical consistency and robustness against stochastic dynamics. Meanwhile, policy rollouts are leveraged as structured informative priors to support the learning phase with Group-Relative Policy Constraint (GRPC) optimizers that impose a group-based adaptive trust-region in the latent action space. This principled combination enables the robot agent to prioritize high-confidence, high-reward behavior while maintaining effective, targeted exploration under uncertainty. Evaluated on the `HumanoidBench` locomotion suite with the Unitree 26-DoF H1-2 humanoid, *DoublyAware* demonstrates improved sample efficiency, accelerated convergence, and enhanced motion feasibility compared to RL baselines. Our simulation results emphasize the significance of structured uncertainty modeling for data-efficient and reliable decision-making in TD-MPC-based humanoid locomotion learning.**

I. Introduction

In model-based reinforcement learning (MBRL) for humanoid locomotion learning, uncertainty is a central concern for ensuring robustness and safe behaviors, particularly for high-dimensional, complex, whole-body coordination, where observations and dynamics can be noisy and stochastic [1], [2]. As humanoid robots need to explore and interact with world dynamics, they must adaptively reason about two fundamentally distinct sources of uncertainty: one inherent to the environment and one arising from limitations in learned policy knowledge. Therefore, in this work, we identify two complementary forms of uncertainty to tackle this problem:

- Planning uncertainty arises from the stochasticity in the local sampling-based trajectory optimizers, such as Model Predictive Control [3], in the planning phase.
- Policy uncertainty stems from the learned policy network's incomplete knowledge due to the unexplored action-state space during the learning phase.

[1]University of Texas at Arlington, Texas, USA
[2]Intelligent Autonomous Systems Lab, TU Darmstadt, Germany
[3]German Research Center for AI (DFKI), SAIROL, Darmstadt, Germany
[4]Hessian.AI, Darmstadt, Germany
[5]Automation & Control Institute (ACIN), TU Wien, Vienna, Austria
[6]Austrian Institute of Technology (AIT) GmbH, Vienna, Austria
E-mails: khang.nguyen8@mavs.uta.edu, minh.vu@ait.ac.at.

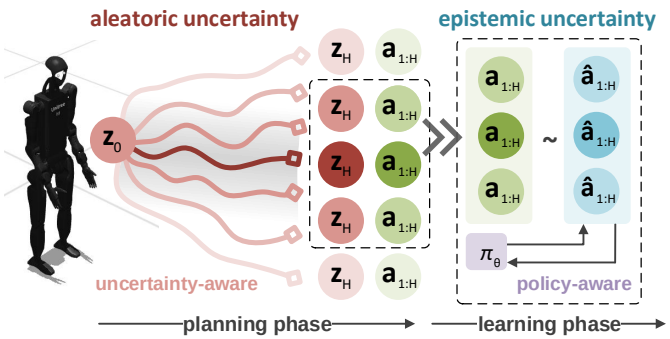

Fig. 1: Overview of *DoublyAware*: Disjoint uncertainty decomposition in TD-MPC frameworks and refinements for each component in planning and learning phases for robust humanoid locomotion.

Planning uncertainty maps to *aleatoric uncertainty*, which is induced by the environment randomness together with the system dynamics (*e.g.*, ground contact, observation noises, and multi-modal nature of feasible movement trajectories). Such uncertainty cannot be eliminated even with augmented data, as it naturally reflects the stochasticity of the world itself. Meanwhile, policy uncertainty is akin to *epistemic uncertainty*, where the model is being trained yet to know about the world due to its internal training experience, which can be reduced through further exploration, learning, and sophisticated policy-aware optimization.

The concept of aleatoric and epistemic uncertainty can be dated back to prior works in machine learning [4], [5]. Vast adaptations have been made for further investigation in learning the world dynamics and control drifts under the influence of uncertainty in classification [4] for feature selection and autonomous driving [6], [7] for trajectory prediction under the influence of control uncertainty and world dynamics. In the scope of learning for control, Temporal Difference Model Predictive Control (TD-MPC) [8], [9] has shown its excellence at short-horizon decision-making through real-time trajectory optimization with TD-learning, enabling more flexible and scalable behavior learning for MBRL-based techniques. Still, these methods frequently suffer from poor sample efficiency and unstable policy updates [10], [11] and notably remain vulnerable to planning compound errors and learning biases issues that are particularly pronounced in high-dimensional control settings [8], [9], [12].

To address this dual challenge, we proposed *DoublyAware*, a planning- and policy-aware TD-MPC-based method for humanoid locomotion learning. Specifically, our approach explicitly decomposes aleatoric planning and epistemic learning uncertainties and solves them distinctively, as illustrated in Fig. 1. Leveraging conformal prediction theory [13], *DoublyAware* integrates conformal quantile filtering to select suitable candidate trajectories robustly, and uses them as informative

priors to modulate agents into a high-reward learning space as a planning-aware solution. Furthermore, inspired by recent advancements in large language models, *DoublyAware* incorporates group-relative policy optimizers [14] in the learning phase with an adaptive trust-region as policy-aware learning. Our contributions are threefold:

1) We outline the decomposition of overall uncertainty in MBRL into planning and policy uncertainty, and solve them distinctively instead of framing them as one.
2) We integrate the planning-aware mechanism with the policy-aware learning optimizer, enabling uncertainty-calibrated trajectory filtering followed by policy rollouts used as informative learnable priors.
3) We evaluate *DoublyAware*'s performance on locomotion tasks in `HumanoidBench` [15], showcasing its improvements compared to baseline methods regarding learning speed and kinodynamically feasible motions.

II. Related Work

Temporal-Difference Model Predictive Control: Humanoid locomotion is one of the most complex control systems, stemming from its high-dimensional continuous action spaces, inherently unstable dynamics, and complex interactions with the environment [16]. TD-MPC has shown promise in addressing these challenges by uniting the short-horizon optimization capabilities of MPC with the sample-efficient, value-driven learning of RL. Prior works on this [8], [9], [17] have demonstrated that incorporating TD learning into MPC frameworks enables flexible value function learning without relying on handcrafted cost functions. Building upon this direction, TD-MPC2 [9] extends the original TD-MPC [8] by introducing scalable latent world models tailored for continuous control to mitigate error accumulation and improve planning stability, where previously even minor errors can quickly lead to destabilized motion as a result. By merging TD learning with MPC-style planning, these frameworks enhance sample efficiency and offer adaptability in high-dimensional control. However, these methods fail when encountering more complex tasks as their uncertainty compounds over time for an extended task execution period. In this work, we further investigate and solve the planning and learning uncertainty distinctively that might lead to poor performance and non-feasible behavior of the TD-MPC framework for humanoid control, especially for locomotion tasks.

Uncertainty-Aware Robot Planning: Recent developments in uncertainty-aware robotics have emphasized the role of conformal prediction and information-theoretic decomposition to enhance planning robustness. In trajectory and motion planning, conformal prediction offers statistical guarantees through distribution-free calibration, making it a natural fit for high-risk, multimodal robotic tasks. Prior works have integrated conformal methods into learned manifold learning [18], [19], enabling model-agnostic risk assessment for learned representations. This research has been extended to motion planning under dynamic uncertainty, where adaptive conformal frameworks improve safety and feasibility [20]–[26]. Other approaches have explored handling distribution shifts during policy learning via conformal mechanisms [27], while others target high-dimensional control for teleoperation through confidence-aware policy mappings [28]. Complementing these, Stochasticity in Motion introduces an entropy-based decomposition of trajectory uncertainty into aleatoric and epistemic terms, formalizing their roles in motion prediction and emphasizing their implications for safe downstream planning. Unlike previous approaches, our work directly applies conformal prediction to latent trajectory selection between policy-guided priors and stochastic trajectories to alleviate exploration uncertainty during planning, offering a solution for effectively tackling the aleatoric part of the TD-MPC framework.

Policy-Aware Optimization for Robot Learning: Policy mismatching poses a core challenge in robot control, particularly for humanoid locomotion, where off-policy methods are highly susceptible to discrepancies between the actions rolled out by the learned policy and the targets generated by temporal-difference updates. The misalignment and bootstrapping error accumulation often lead to compounding inaccuracies and poor generalization performance [29], [30]. Offline RL approaches have significantly addressed this by learning from fixed datasets. Several methods mitigate distributional shift by explicitly regularizing the policy toward expert demonstrations [29], [31], while others leverage importance sampling to correct for distributional mismatch in value estimation [32], [33]. Similarly, in-sample learning techniques [34], [35] avoid out-of-distribution actions by constraining updates to observed data, implicitly ensuring policy reliability. This challenge is equally framed in MBRL; for example, LOOP [17] introduces actor regularization to inject conservatism into the planning process and stabilize learning. Departing from prior work, our method enforces distributional consistency directly on the policy prior in latent space, without modifying the underlying planner, allowing for more flexible planning while maintaining stability and enabling fast policy adaptation. In brief, our work extends these principles by incorporating algorithmic stability and data efficiency through Group-Relative Policy Optimization (GRPO) [14] with an explicit trust-region constraint for policy optimization during the learning phase.

III. Planning- & Policy-Aware Temporal Difference Learning

A. Vanilla Planning with Model-Predictive Control

Humanoid locomotion tasks can be formulated as infinite-horizon Markov Decision Processes (MDPs), defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \rho, r, \gamma)$, where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ is the action space, $\rho : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ represents the transition function, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, and $\gamma \in (0, 1]$ is the discount factor. The objective J^π is to learn the parameters θ for the policy network $\Pi_\theta : \mathcal{S} \to \mathcal{A}$ that enables the robot to continuously generate optimal actions, maximizing the expected discounted cumulative reward over a trajectory, ξ, as a sequence of states and actions:

$$J^\pi = \mathbb{E}_{\xi \sim \Pi_\theta} \left[\sum_{t=0}^{\infty} \gamma^t r(\mathbf{s}_t, \mathbf{a}_t) \right], \; \xi = \begin{bmatrix} \mathbf{s}_0 & \mathbf{s}_1 & \dots \\ \mathbf{a}_0 & \mathbf{a}_1 & \dots \end{bmatrix} \quad (1)$$

with each action $\mathbf{a}_t$ is sampled from the policy $\Pi_\theta(\mathbf{s}_t)$, and each subsequent state $\mathbf{s}_t$ is determined by $\rho(\mathbf{s}_{t-1}, \mathbf{a}_{t-1})$ based on the previous state $\mathbf{s}_{t-1}$ and action $\mathbf{a}_{t-1}$.

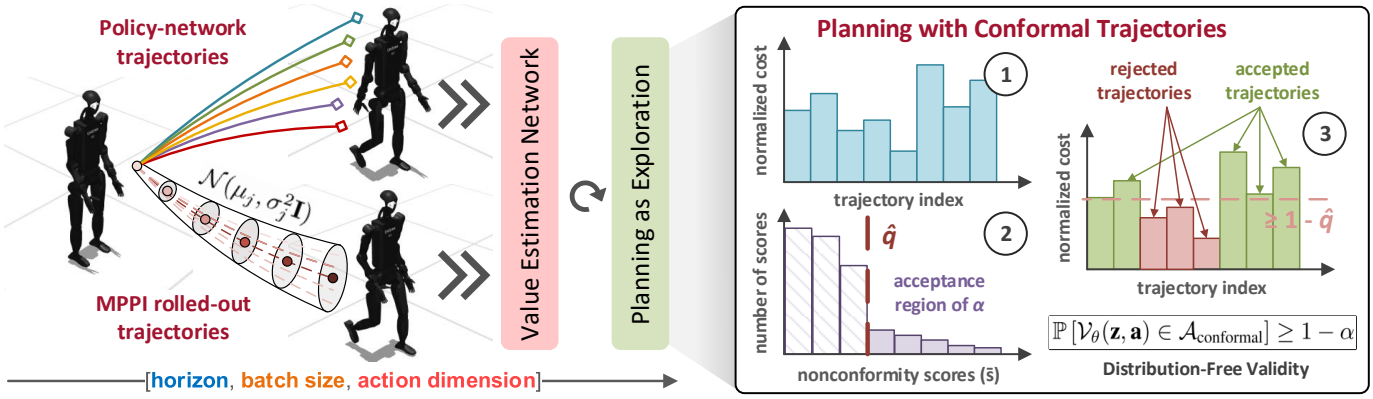

Fig. 2: Uncertainty-Aware Planning for Humanoid Locomotion: At each planning step, two sets of trajectories are sampled from the policy network and MPPI planner. The policy network provides prior-guided trajectories that reflect learned behavior, while MPPI samples explore around a sampling distribution $\mathcal{N}(\mu_j, \sigma_j^2\mathbf{I})$ with j as the number of iterations. Each iteration contains a batch of trajectories for all action dimensions along the predictive planning horizon. (1) These candidate trajectories are then evaluated using the TD-based cost function, (2) assigned normalized nonconformity scores $\bar{s}$, (3) and filtered via a conformal quantile threshold $\hat{q}$ to retain only trajectories within the empirical $(1-\alpha)$ prediction set. These conformal latent trajectories ensure statistically reliable planning under policy model error and guarantee quality explorative behavior for sequential decision-making while learning humanoid locomotion tasks.

Hansen *et al.* [8], [9] introduced an approach that integrates MPPI [36] as a local trajectory optimizer for short-horizon planning within a learned latent dynamics model. During the planning phase, the action sequences of length H are sampled as latent trajectories from the learned dynamics model, and the cumulative return ϕ_ξ of each sampled trajectory ξ is calculated using a trained value estimator $\mathcal{V}_\theta(\mathbf{z}_t, \mathbf{a}_t)$, as the composition of $R_\theta(\mathbf{z}_t, \mathbf{a}_t)$ and $Q_\theta(\mathbf{z}_H, \mathbf{a}_H)$ as follows:

$$\mathcal{V}_\theta(\mathbf{z}_t, \mathbf{a}_t) \coloneqq \phi_\xi = \sum_{t=0}^{H-1} \gamma^t R_\theta(\mathbf{z}_t, \mathbf{a}_t) + \gamma^H Q_\theta(\mathbf{z}_H, \mathbf{a}_H), \quad (2)$$

where $\mathbf{z}_t = h_\theta(\mathbf{s}_t)$ is the latent representation that selectively captures the relevant dynamics of the state $\mathbf{s}_t$, rather than all observation dimensions, $\mathbf{z}_t = d_\theta(\mathbf{s}_{t-1}, \mathbf{a}_{t-1})$ represents the next latent state under the latent dynamics d_θ, $\hat{r}_t = R_\theta(\mathbf{s}_t, \mathbf{a}_t)$ and $\hat{q}_t = Q_\theta(\mathbf{s}_t, \mathbf{a}_t)$ denote reward and value estimators, and $\mathbf{a}_t \sim \mathcal{N}(\mu_t, \sigma_t^2\mathbf{I})$ describes the MPPI sampling distributions:

$$\mu_j = \eta \sum_{i=1}^{k} \Omega_i \xi_i^*, \quad \sigma_j^2 = \eta \sum_{i=1}^{k} \Omega_i (\xi_i^* - \mu^j)^2, \quad (3)$$

where $\Omega_i = \exp(\tau \phi_{\xi,i}^*)$, τ is a temperature parameter, η represents the normalizing term with respect to the coefficients Ω_i, and ξ_i^* denotes the i^{th} of the latent trajectory corresponding to return estimate ϕ_ξ^*. In the MBRL-based robot learning, Eq. 2 can therefore be called as an H-step look-ahead policy, which iteratively maximizes the first step's costs.

B. Planning as Exploration with Conformal Trajectories

To improve planning exploration probabilistically safely, we integrate conformal prediction into the MPPI planner, as depicted in Alg. 1. Specifically, conformal prediction allows the underlying planner to choose "*statistically-significant*" trajectories as candidates without assuming any parametric form of the return value/cost distribution.

The planning procedure in TD-MPC2 [9] incorporates two distinct sources of candidate latent trajectories (Fig. 2):

- The first set is sampled from the current policy network by simulating trajectories under the learned world model, which acts as prior knowledge by proposing trajectories that reflect the agent's learned behavior so far. These trajectories serve as a warm start for planning, reducing dependence on purely random sampling and encouraging consistency during training.
- The second set is generated via the MPPI planner, which aims to explore the action space heuristically and re-weights them based on expected cumulative rewards, thus refining the action distribution toward higher-value trajectories through the learning process.

Using policy-guided and MPPI samples improves exploration and stability by balancing prior-driven guidance with adaptive search. Nevertheless, both sources might fall into sub-optimality, where policy rollouts may propagate error-prone behavior during training, while MPPI may overfit to inaccuracies in the model dynamics. Thus, an uncertainty-aware selection mechanism for the MPPI planner is needed to calibrate and filter trajectories based on empirical plausibility.

We compute the conformal scores to quantify the agreement between candidate trajectories and the prior trajectories from the policy network. Denote $\mathcal{A}_\pi = \{\mathbf{a}_{1:H}^{\pi,(i)}\}_{i=1}^{N_\pi}$ as the N_π prior trajectories sampled from the policy with $v_{1:H}^{\pi,(i)}$ are their evaluated costs, $\mathcal{A}_{\text{mppi}} = \{\mathbf{a}_{1:H}^{(i)}\}_{i=1}^{N}$ as the N candidate trajectories generated via MPPI with $v_{1:H}^{(i)}$ as their corresponding costs. With Eq. 2, the nonconformity scores can be directly computed from the latent trajectories' costs:

$$\bar{s}^{(i)} = 1 - \eta \mathcal{V}_\theta \left(\mathbf{z}_{1:H}^{(i)}, \mathbf{a}_{1:H}^{(i)} \right) \ \forall \mathbf{a} \in \mathcal{A} = \{\mathcal{A}_\pi \cup \mathcal{A}_{\text{mppi}}\} \quad (4)$$

where the scores are computed as the normalized cost functions across all actions $\mathbf{a}_{1:H}^{(i)}$ along the planning horizon H. Note that normalized costs are computed to align with the original concept of softmaxes, in the range of 0 to 1, in classical conformal classification problems.

As both sets are exchangeable and the TD-based cost function $\mathcal{V}_\theta(\mathbf{z}, \mathbf{a})$ is used to estimate the conformal scores, the conformal prediction set $\mathcal{A}_{\text{conformal}}$ satisfies the marginal

Alg. 1: Planning with Conformal Trajectories

Input: H: planning horizon, J: number of iterations, $\mathcal{N}(\mu_0, \sigma_0^2\mathbf{I})$: initial distribution for MPPI, $\mathbf{z}_t$: latent represent. at time t, α: error rate, N_π: number of prior trajectories by π_θ, N: number of MPPI-sampled trajectories

Output: $\mathbf{a}_t$: action sampled from $\mathcal{N}(\mu_J, \sigma_J^2\mathbf{I})$

function plan(H, J, $\mathbf{z}_t$, α, $\mathcal{N}(\mu_0, \sigma_0^2\mathbf{I})$, N, N_π)
  while *planning* **do**
    for $j = 1, \ldots, J$ **do**
      $\mathcal{A}_\pi \leftarrow \{\mathbf{a}_i^\pi\}_{i=1}^{N_\pi} \sim \pi_\theta$
      $\mathcal{A}_{\text{mppi}} \leftarrow \{\mathbf{a}_i^{\text{mppi}}\}_{i=1}^{N} \sim \mathcal{N}(\mu_{j-1}, \sigma_{j-1}^2\mathbf{I})$
      for $\mathbf{a}_i \in \mathcal{A} = \{\mathcal{A}_\pi \cup \mathcal{A}_{mppi}\}$ **do**
        $v_i \leftarrow 0$, $\mathbf{z}_0 \leftarrow \mathbf{z}_t$
        for $t = 0, \ldots, H-1$ **do**
          $v_i \leftarrow v_i + \gamma^t R_\theta(\mathbf{z}_t, \mathbf{a}_{i,t})$
          $\mathbf{z}_{t+1} \leftarrow d_\theta(\mathbf{z}_t, \mathbf{a}_{i,t})$
        $v_i \leftarrow v_i + \gamma^H Q_\theta(\mathbf{z}_H, \mathbf{a}_{i,H})$ (Eq. 2)
      // compute nonconformity scores from costs
      $\bar{s}_i \leftarrow 1 - \texttt{normalize}(v_i)$ (Eq. 4)
      $\hat{q} \leftarrow \texttt{quantile}(\{\bar{s}_i\}, 1-\alpha)$ (Eq. 6)
      $\mathcal{A}_{\text{conformal}} \leftarrow \{i \in \mathcal{A} \mid \bar{s}_i \le \hat{q}\}$ (Eq. 7)
      // update parameters for next planning step
      $\mu_j, \sigma_j \leftarrow \texttt{update}(\{\mathbf{a}_i\}_{i \in \mathcal{A}_{\text{conformal}}})$ (Eq. 3)
  return $a_\tau \sim N(\mu, \sigma^2\mathbf{I})$ **for** $\tau = t, \ldots, t+H$

coverage guarantee, so-called distribution-free validity:

$$1 - \alpha + \tfrac{1}{(N_\pi + N)+1} \ge \mathbb{P}\left[\mathcal{V}_\theta(\mathbf{z}, \mathbf{a}) \in \mathcal{A}_{\text{conformal}}\right] \ge 1 - \alpha. \quad (5)$$

With the nonconformity scores $\bar{s}^{(i)}$ from Eq. 4, we construct the prediction set of candidate trajectories that conform to the statistical properties of the calibration set. The key idea is to select a quantile threshold $\hat{q}$ such that a fixed proportion $(1-\alpha)$ of calibration trajectories achieve scores less than or equal to this threshold. Formally, we define it:

$$\hat{q} = \texttt{quantile}\left[\left\{\bar{s}^{(i)}\right\}_{i=1}^{N_\pi + N}, (1-\alpha)\right], \quad (6)$$

where $\alpha \in [0, 1)$ is the pre-defined risk level that controls the desired coverage. Intuitively, $\hat{q}$ defines a level set of conformity: any candidate MPPI latent trajectory with a score $\bar{s}^{(j)} \le \hat{q}$ is deemed statistically compatible with the teacher prior $\mathcal{A}_\pi$. This leads to the formal definition of the conformal prediction set that is in Eq. 5, described as:

$$\mathcal{A}_{\text{conformal}} = \left\{\mathbf{a}_{1:H}^{(j)} \in \mathcal{A} \mid \bar{s}^{(j)} \le \hat{q}\right\}, \quad (7)$$

which serves as a filtered subset of high-confidence trajectories. The conformal set $\mathcal{A}_{\text{conformal}}$ admits finite-sample validity guarantees under the assumption that the calibration scores $\{\bar{s}^{(i)}\}$ and the candidate scores $\{\bar{s}^{(j)}\}$ are exchangeable for any i and j within the union set size. Thus, $\mathcal{A}_{\text{conformal}}$ contains the best trajectories with probability at least $1-\alpha$, which satisfies Eq. 5. Moreover, in online learning, the value estimator is improved over time as it learns periodically.

For this reason, we suppose that the TD-based cost function $\mathcal{V}_\theta(\mathbf{z}, \mathbf{a})$ is an imperfectly taught value estimator (*i.e.*, a weak teacher in conformal prediction theory [13]) that reveals informative conformity scores only along a teaching schedule $\mathcal{L} = \{n_k\}_{k \ge 1} \subset \mathbb{N}$. If the teaching schedule satisfies the sub-linearity condition $\lim_{k\to\infty}(n_k/n_{k-1}) = 1$, the conformal quantile $\hat{q}$ in Eq. 6, constructed from calibration scores $\{\bar{s}^{(i)}\}$, admits asymptotic weak validity. That is, the prediction set $\mathcal{A}_{\text{conformal}}$ defined in Eq. 7 satisfies:

$$\liminf_{\mathcal{L}\to\infty} \mathbb{P}\left[\bar{s}^{\text{test}} \le \hat{q}\right] \ge 1 - \alpha. \quad (8)$$

This means that this asymptotic validity under weak priors (Eq. 8) holds even if the cost signal from $\mathcal{V}_\theta(\mathbf{z}, \mathbf{a})$ is imperfect or inconsistent, as long as the prior model improves.

Indeed, Eq. 2 is a black-box oracle, and no distributional assumptions are made beyond exchangeability within this context. Therefore, this conformal filtering preserves the model-agnostic nature while seamlessly integrating with latent dynamics and value-based planning, highlighting that the delay in learning the value estimator is allowed.

C. Learning with Group Relative Policy Constraint

We adopt and improve GRPO [14] to enhance action group-based explicit advantage estimation, enhancing entropy-regularized policy gradient methods by leveraging group-wise action comparisons, enabling the policy to learn from relative action preferences rather than relying on those priors. In standard actor-critic methods, policy gradients are scaled by absolute values or advantage estimates, which may be sensitive to estimation errors. These limitations become especially pronounced in long-horizon tasks; GRPO addresses this issue by constructing a relative preference distribution across sampled groups. Mathematically, at each state $\mathbf{s}_i$, a set of G actions $\{\mathbf{a}_i^1, \ldots, \mathbf{a}_i^G\}$ is sampled, and their Q-values $\{q_i^1, \ldots, q_i^G\}$ are computed, which are used to compute the group-based soft attention advantage scores:

$$A_i(\mathbf{q}) = \frac{\exp(q_i/\tau)}{\sum_{g=1}^{G} \exp(q_i^g/\tau)}, \quad (9)$$

where $\mathbf{q} = Q_\theta(\mathbf{s}, \mathbf{a}_i)$ denotes the estimates, and τ is a temperature parameter and $0 \le A_i(\cdot) \le 1$ as its property.

As $\{\mathbf{a}_i^g\}_{g=1}^G$ is a group of G actions sampled from a policy $\pi_\theta(\mathbf{s})$ at the state $\mathbf{s}$ with $\hat{r}_i = r_\theta(\mathbf{s}, \mathbf{a}_i)$ and $\hat{q}_i = Q_\theta(\mathbf{s}, \mathbf{a}_i)$ are the reward and estimated values, respectively, we assume that $||\nabla_\theta \log \pi_\theta(\mathbf{a} \mid \mathbf{s})|| = C$, and $\hat{q}_i$, $\hat{r}_i$ are bounded above with $\forall \mathbf{a} \in \mathcal{A}$ and $\forall \mathbf{s} \in \mathcal{S}$. Therefore, we obtain:

$$\text{Var}\left[\nabla_\theta \mathcal{L}_{\text{softmax}}\right] \le \text{Var}\left[\nabla_\theta \mathcal{L}_{\text{std-norm}}\right], \quad (10)$$

with $\mathcal{L}_{\text{softmax}}$ and $\mathcal{L}_{\text{std-norm}}$ are the softmax-based and standard normalized advantage losses, respectively. The variance of gradient of $\mathcal{L}_{\text{softmax}}$ is thus smaller than that of $\mathcal{L}_{\text{std-norm}}$:

$$||\nabla_\theta \mathcal{L}_{\text{softmax}}|| \text{ is bounded, } ||\nabla_\theta \mathcal{L}_{\text{std-norm}}|| \text{ is unbounded} \quad (11)$$

yields more stable policy updates at some constant C that asymptotically bounds $||\nabla_\theta \log \pi_\theta(\mathbf{a} \mid \mathbf{s})||$. Two keys favor softmax-based over normalized advantages. First, their outputs lie between 0 and 1, limiting the impact of outliers. Meanwhile, normalized advantages induce large magnitudes under noise, leading to high-variance gradients. Second, policy gradients scale with the advantage values. The gradient steps

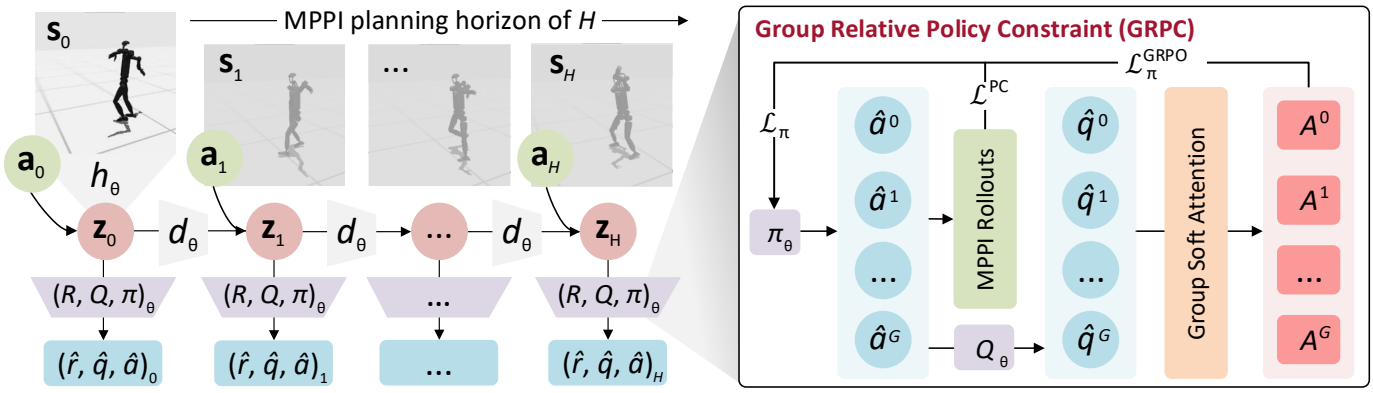

Fig. 3: Policy-Aware Learning for Humanoid Locomotion: Given an initial state $\mathbf{s}_0$, it is encoded into a latent representation $\mathbf{z}_0$ using the encoder h_θ. A latent dynamics model d_θ then iteratively predicts future latent states $\mathbf{z}_{t+1}$ over a planning horizon of H steps, conditioned on actions $\mathbf{a}_t$ and current latent states $\mathbf{z}_t$. Throughout this horizon, reward estimates, Q-values, and policy outputs are generated via learned networks R_θ, Q_θ, and π_θ, respectively, supporting trajectory optimization in latent space guided by temporal-difference objectives. For each state, groups of sampled actions are rolled out and evaluated through Q-values, which are transformed into softmax-weighted advantage scores A^g per action group g. These scores inform the GRPC objective, promoting the selection of high-value actions while reducing policy variance. A trust-region is enforced through a KL divergence penalty between the current policy and its MPPI-based prior, thereby regularizing residual policy learning and bounding divergence from prior behavior to ensure stable updates.

will be disproportionately unstable if the advantage is large or small. Therefore, softmax-based advantages smooth out extreme values and act like a soft attention mechanism, giving more stable policy updates during training episodes.

With the group relative weights in Eq. 9 and based on Eq. 10 and Eq. 11, the improved GRPO objective is defined as:

$$\mathcal{L}_\pi^{\text{GRPO}}(\theta) = \frac{1}{G}\sum_{i=1}^{G} A_i(\mathbf{q}) \log \pi_\theta(\mathbf{a}_i \mid \mathbf{s}) \tag{12}$$

where μ_k denotes the behavior policy at k^{th} iteration from $\mathcal{D}$ in Eq. 3. The KL constraint ensures the updated policy remains within a trust region of π. The overall policy objective combines the trust-region matching with Eq. 12:

$$\mathcal{L}_\pi(\theta) = \frac{1}{G}\sum_{i=1}^{G} A_i(\mathbf{q}) \log \pi_\theta(\mathbf{a}_i \mid \mathbf{s})) + \beta \log \mu(\mathbf{a} \mid \mathbf{s}), \tag{13}$$

where β is a weighting coefficient controlling the penalty strength, the second term of Eq. 13 imposes a residual-style regularization as a trust-region for policy optimization [37].

Meanwhile, the latent dynamics d_θ, encoder h_θ, reward network R_θ, and value network Q_θ are concurrently optimized by the following model objective:

$$\mathcal{L}(\theta;\ \xi_i) = \|d_\theta(\mathbf{z}_i, \mathbf{a}_i) - h_\theta(\mathbf{s}_{i+1})\|_2^2 \tag{14a}$$
$$+ \|R_\theta(\mathbf{z}_i, \mathbf{a}_i) - r_i\|_2^2 \tag{14b}$$
$$+ \|Q_\theta(\mathbf{z}_i, \mathbf{a}_i) - [r_i + \gamma Q_\theta(\mathbf{z}_{i+1}, \pi_\theta(\mathbf{z}_{i+1}))]\|_2^2 \tag{14c}$$

The training procedure with TD learning with GRPC at each short horizon for long-horizon locomotion tasks is summarized in Alg. 2 and is described visually in Fig. 3.

Alg. 2: Learning with Group-Relative Constraint

Input : T : trajectory length, H: planning horizon, G: number of groups, $\mathcal{D}$: latent buffer, S: number of iterations

function `learn`$(T,\ H,\ G,\ \mathcal{D},\ S)$
  while *learning* **do**
    for $t = 0, \ldots, T$ **do**
      $\mathbf{a}_t \sim \Pi_\theta(h_\theta(\mathbf{s}_t))$
      $(\mathbf{s}_{t+1}, r_t) \sim \mathcal{P}(\mathbf{s}_t, \mathbf{a}_t), \mathcal{R}_\theta(\mathbf{s}_t, \mathbf{a}_t)$
      $\mathcal{D} \leftarrow \mathcal{D} \cup (\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1})$
    for *step = 0, …, S* **do**
      $\{\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1}\}_{t:t+H}^{g} \sim \mathcal{D}$ **for** G groups
      $\mu^G, \sigma^G =$ `compute_moments`$(\mathbf{a}_t)$ (Eq. 3)
      $\mathbf{z}_t \leftarrow h_\theta(\mathbf{s}_t)$ **if** $\mathbf{s}_t$ is the first observation
      for $i = t, \ldots, t+H$ **do**
        $\mathbf{z}_{i+1} \leftarrow d_\theta(\mathbf{z}_i, \mathbf{a}_i)$ (Eq. 14a)
        $\hat{r}_i \leftarrow R_\theta(\mathbf{z}_i, \mathbf{a}_i)$ (Eq. 14b)
        // group sampling & policy constraint
        for $g = 1, \ldots, G$ **do**
          $\widehat{\mathbf{a}}_i^g \sim \pi_\theta(\mathbf{z}_i)$
          $\varepsilon = (\widehat{\mathbf{a}}_i^g - \mu^G)/\sigma^G$
          $\widehat{\mathbf{a}}_i^g \leftarrow$ `threshold`$(\widehat{\mathbf{a}}_i^g, \varepsilon)$
          $\hat{q}_i^g = Q_\theta(\mathbf{z}_i, \widehat{\mathbf{a}}_i^g)$ (Eq. 14c)
        $A_i^g =$ `softmax`$(\hat{\mathbf{q}}^g)$ (Eq. 9)
        $\mathcal{L}_\pi^{(i)} = \frac{1}{G}\sum_{g=1}^{G} A_i^g \log \pi_\theta(\widehat{\mathbf{a}}_i^g \mid \mathbf{z}_i)$
      $\mathcal{L}_\pi = \frac{1}{H}\sum_{i=t}^{t+H}\left[\mathcal{L}_\pi^{(i)} + \beta \mathcal{L}_{\text{KL}}\right]$ (Eq. 13)
      $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_\pi$

IV. Simulation Results & Ablation Studies

To assess our proposed method's performance, we train *DoublyAware* on **the Unitree 26-DoF H1-2 humanoid**, comprised of two legs of 12 DoFs and two arms of 14 DoFs. The torso joint is locked to eliminate unnecessary body-turning actions. The hands are included in the robot's model during training to account for their mass, ensuring that the learned policies take both hands into consideration, even though the robot is not involved in manipulation tasks. We evaluate *DoublyAware*'s performance against SAC [38], BC-SAC [39], AWAC [40], TD-MPC2 [9], TD-M(PC)^2 [12] on the locomotion tasks in `HumanoidBench` [15]. Nine tasks include standing, walking, running, sitting on a chair, crawling through a tunnel, navigating through standing poles, hurdling,

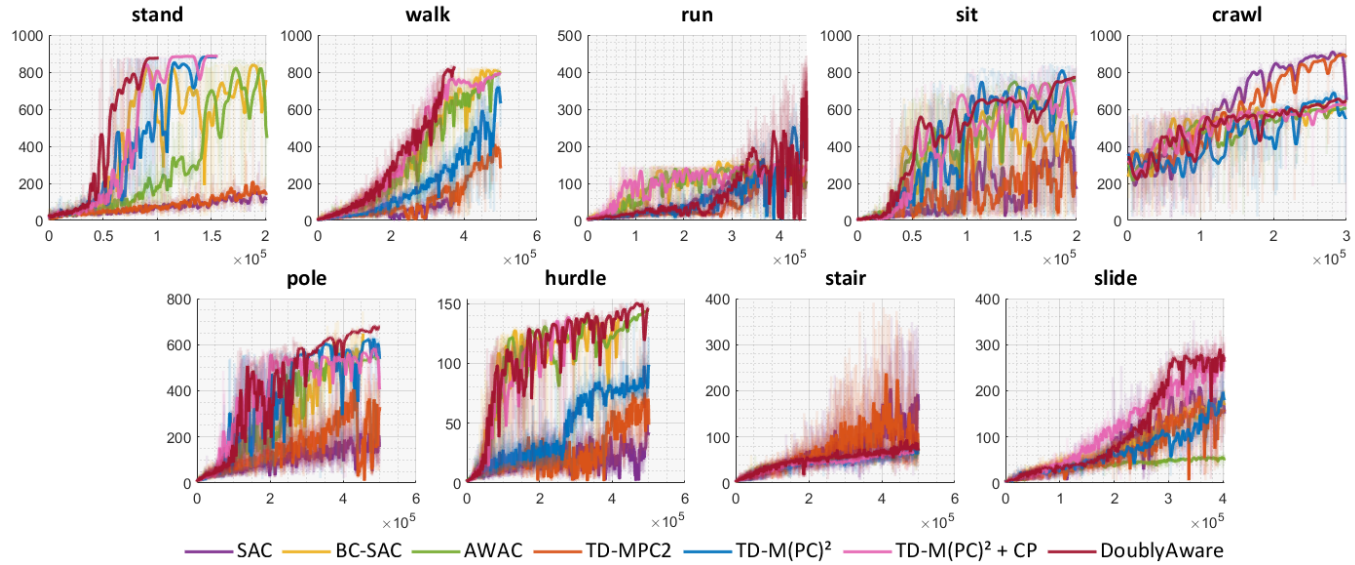

Fig. 4: Episode Returns of *DoublyAware* and Baselines on H1–2 in Locomotion Tasks: *DoublyAware* achieves rapid convergence over others in standing, walking, sitting, navigating through poles, hurdling, and sliding tasks, while it performs worse in more complex tasks such as crawling and stair-climbing, which require high-dimensional whole-body coordination. In general, *DoublyAware* shows slightly better data-efficiency than TD-M(PC)^2, BC-SAC, and AWAC and significantly better sampling-efficiency than SAC and TD-MPC2.

stair-climbing, and walking over slides. Our evaluations are to answer the following questions:

1) Does *DoublyAware* achieve superior reward convergence compared to existing methods?
2) Can *DoublyAware* successfully solve tasks that demand long-horizon, whole-body coordination?
3) Whether conformal trajectory planning and group-relative policy constraint learning mutually benefit?
4) How does the motion generated by *DoublyAware* compare qualitatively to those from other baselines?

Across the experiments, the evaluated algorithms are set with the default starting pose of the H1-2. The hyperparameters include the planning horizon of 3, batch size of 256, action dimension of 26, learning rate of 0.0003, and number of prior trajectories of 24 on an AMD Ryzen 9 7950X3D CPU and an NVIDIA RTX 4090 GPU. Additionally, we use a group number of 3 for group-based policy optimization and an error rate of 0.05 for conformal trajectory planning.

A. Episode Returns of Locomotion Tasks

We report the episode return comparisons across all evaluated methods on the `HumanoidBench` locomotion tasks in Fig. 4. On foundational tasks such as standing, walking, running, and sitting, *DoublyAware* demonstrates significantly faster convergence than competing approaches. Specifically, the H1-2 humanoid achieves upright standing in fewer than $100,000$ training iterations, walking in approximately $300,000$ iterations, running in $450,000$ iterations, and sitting in $200,000$ iterations. In contrast, baseline methods such as SAC, BC-SAC, AWAC, TD-MPC2, and TD-M(PC)^2 show their difficulties achieving comparable performance on walking and running within the same training budget.

On more complex tasks requiring precise whole-body coordination, such as navigating around standing poles, hurdling over obstacles, and walking across inclined slides, *DoublyAware* maintains a consistent performance advantage. Notably, it surpasses a reward threshold of 600 on the pole navigation task within $500,000$ iterations, achieves a reward of 150 on hurdling, and reaches around 300 on walking over slides. While other methods eventually learn these tasks, they exhibit slower convergence rates and higher variability in performance. On the most challenging tasks, such as crawling and stair-climbing, *DoublyAware* underperforms relative to other methods in terms of reward acquisition within the same number of training iterations. Overall, *DoublyAware* shows the training efficiency compared to the competing baselines.

B. Ablation Studies on Uncertainty-Aware Modules

As also shown in Fig. 4, we conduct an ablation analysis to examine whether conformal trajectory prediction (CP) and group-relative policy constraint (GRPC) offer complementary benefits when integrated into the TD-MPC framework. We compare three variants: TD-M(PC)^2 (blue) – which is plain baseline, TD-M(PC)^2 + CP (pink), and *DoublyAware* (red) – which combines both CP and GRPC.

Across most locomotion tasks, *DoublyAware* consistently outperforms the ablated variants in final performance and sample efficiency. For example, in tasks such as standing, walking, navigating through poles, hurdling, and walking over slides, *DoublyAware* reaches peak returns faster and more stably than the other two. Although it underperforms in complex coordination tasks like crawling and stair-climbing, TD-M(PC)^2 + CP performs better in these settings, benefiting from CP's uncertainty-aware planning – yet it lacks generalization across broader task domains. In general, these results suggest that CP and GRPC are mutually beneficial, each addressing complementary aspects of the problem: CP improves robustness and uncertainty-awareness in trajectory exploration, and GRPC enhances policy-aware policy optimization for learning. Their integration in *DoublyAware* leads to a more data-efficient and robust locomotion policy.

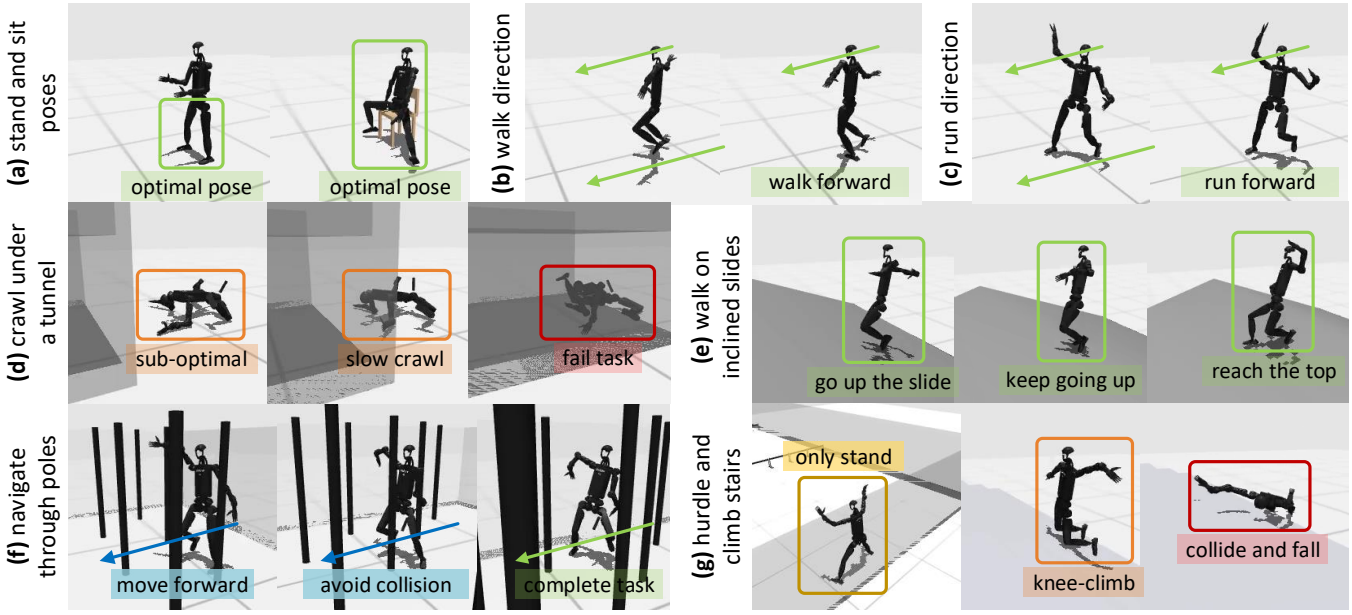

Fig. 5: Qualitative Results of H1-2 in Locomotion Tasks: (a) *Standing & Sitting Poses: DoublyAware* enables H1-2 to complete these tasks with appropriate leg and hand poses. (b) and (c) *Walking & Running Direction:* H1-2 can walk and run in the forward direction when trained with *DoublyAware*, unlike when being trained with other algorithms: TD-M(PC)^2 induces walking/running backward, others generate dynamically-infeasible motions. (d) *Crawling under A Tunnel: DoublyAware* unable to teach the robot to crawl through the tunnel, where TD-M(PC)^2 and the ablation methods also fail. Only SAC and TD-MPC2 can accomplish this task. (e) *Walking on Inclined Slides:* The H1-2 can reach the top of the slide-like hill faster than other baselines when learning with *DoublyAware*. (f) *Navigating through Poles:* Again, *DoublyAware* generates feasible motions for the robot to move forward and avoid collision; other methods fail on this task. and (g) *Hurdling & Stair-Climbing: DoublyAware* is unable to teach the H1-2 to accomplish these challenging tasks, so do other methods.

C. Visualization of Sequential Behaviors

Beyond task completion, we study the qualitative results of generated behavior across learning algorithms. Fig. 5 shows the performance of H1-2 in a range of locomotion tasks.

In Fig. 5a, *DoublyAware* enables the robot to achieve stable and plausible poses for standing and sitting, demonstrating coordinated control of the legs and limbs. In contrast, baseline methods exhibit unstable or implausible configurations. Fig. 5b and Fig. 5c both show that *DoublyAware* consistently results in coherent walking and running in the forward direction, while TD-M(PC)^2 and other methods induce backward locomotion. For Fig. 5e, in the slide-walking task, *DoublyAware* guides the robot to ascend the incline progressively and reach the top.

In the more challenging scenarios, such as Fig. 5d, crawling through a tunnel, *DoublyAware* fails to complete the task, showing sub-optimal postures and stalled progress, similar to baseline methods, which also fail to solve this task. Only SAC and TD-MPC2 can complete this task sufficiently. Fig. 5f shows that *DoublyAware* improves spatial awareness in navigating between standing poles, generating trajectories that avoid collisions while maintaining forward progress, unlike competing methods that get stuck or misstep. Lastly, in Fig. 5g, DoublyAware and also other baselines fail to teach the H1-2 hurdle and climb stairs. In general, these results highlight that while *DoublyAware* significantly improves performance on many tasks compared to its competitors.

We summarize solving abilities across all locomotion tasks for all algorithms in Table I. Based on both quantitative and qualitative results analyzed, the summary shows an empirical improvement when training the H1-2 with *DoublyAware*. In specific, *DoublyAware* successfully solves most locomotion tasks, including standing, walking, running, sitting on a chair, navigating through standing poles, and walking on an inclined slide. However, *DoublyAware* and others cannot solve crawl, hurdle, and climb stairs tasks, which require more advanced whole-body coordination and dynamic skill refinement. Over-

TABLE I: Solving ability of SAC [38], BC-SAC [39], AWAC [40], TD-MPC2 [9], TD-M(PC)^2 [12], and our method, *DoublyAware*, of locomotion tasks on H1-2 in `HumanoidBench` [15]: ✓ for tasks that are solved sufficiently, ● for tasks that need additional mild refinements for success, and ✗ for tasks that need further intensive learning of whole-body and selective dynamics features.

Method / Task	`stand`	`walk`	`run`	`sit on chair`	`crawl under tunnel`	`navigate through poles`	`hurdle`	`climb stairs`	`walk on inclined slide`
SAC [38]	✗	✗	✗	✗	✓	✗	✗	✗	●
BC-SAC [39]	✗	✗	✗	●	●	●	✗	✗	✗
AWAC [40]	●	✗	✗	●	●	●	✗	✗	✗
TD-MPC2 [9]	✗	✗	✗	●	✓	✗	✗	✗	●
TD-M(PC)^2 [12]	✓	●	●	●	●	●	✗	✗	●
DoublyAware	✓	✓	✓	✓	●	✓	✗	✗	✓

all, these results highlight the robustness of *DoublyAware* in diverse locomotion settings, with room for improvement in tasks demanding complex full-body movements. For more comprehensive results, the demonstration video can be seen at: `https://www.acin.tuwien.ac.at/f3c8/`.

V. Conclusions

This work presents *DoublyAware*, an uncertainty-aware extension of TD-MPC tailored for robust and sample-efficient humanoid locomotion. By decomposing uncertainty into disjoint planning and policy components, our method enables principled reasoning and mitigation in planning and learning phases. Conformal quantile filtering ensures statistically grounded trajectory selection under aleatoric uncertainty, while GRPC with adaptive trust-region regularization promotes stable and policy-aware learning under epistemic uncertainty. Our evaluations on `HumanoidBench` demonstrate that *DoublyAware* surpasses prior methods in convergence speed and motion quality across various whole-body locomotion tasks, highlighting the benefits of structured uncertainty modeling in complex, high-dimensional humanoid control settings.

References

[1] I. Radosavovic, T. Xiao, B. Zhang, T. Darrell, J. Malik, and K. Sreenath, "Learning humanoid locomotion with transformers," *CoRR*, 2023.

[2] ——, "Real-world humanoid locomotion with reinforcement learning," *Science Robotics*, vol. 9, no. 89, p. eadi9579, 2024.

[3] B. Kouvaritakis and M. Cannon, "Model predictive control," *Switzerland: Springer International Publishing*, vol. 38, no. 13-56, p. 7, 2016.

[4] M. H. Shaker and E. Hüllermeier, "Aleatoric and epistemic uncertainty with random forests," in *International Symposium on Intelligent Data Analysis*. Springer, 2020, pp. 444–456.

[5] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods," *Machine learning*, vol. 110, no. 3, pp. 457–506, 2021.

[6] A. Distelzweig, A. Look, E. Kosman, F. Janjoš, J. Wagner, and A. Valada, "Stochasticity in motion: An information-theoretic approach to trajectory prediction," *arXiv preprint arXiv:2410.01628*, 2024.

[7] S. Hagedorn, A. Distelzweig, M. Hallgarten, and A. P. Condurache, "Learning through retrospection: Improving trajectory prediction for automated driving with error feedback," *arXiv preprint arXiv:2504.13785*, 2025.

[8] N. A. Hansen, H. Su, and X. Wang, "Temporal difference learning for model predictive control," in *International Conference on Machine Learning*. PMLR, 2022, pp. 8387–8406.

[9] N. Hansen, H. Su, and X. Wang, "Td-mpc2: Scalable, robust world models for continuous control," in *The Twelfth International Conference on Learning Representations*.

[10] K. Chua, R. Calandra, R. McAllister, and S. Levine, "Deep reinforcement learning in a handful of trials using probabilistic dynamics models," *NeurIPS*, 2018.

[11] A. Argenson and G. Dulac-Arnold, "Model-based offline planning," *arXiv preprint arXiv:2008.05556*, 2020.

[12] H. Lin, P. Wang, J. Schneider, and G. Shi, "Improving td-mpc through policy constraint," *arXiv preprint arXiv:2502.03550*, 2025.

[13] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic learning in a random world*. Springer, 2005, vol. 29.

[14] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu *et al.*, "Pushing the limits of mathematical reasoning in open language models," *arXiv preprint arXiv:2402.03300*, 2024.

[15] C. Sferrazza, D.-M. Huang, X. Lin, Y. Lee, and P. Abbeel, "Humanoidbench: Simulated humanoid benchmark for whole-body locomotion and manipulation," *arXiv preprint arXiv:2403.10506*, 2024.

[16] J. Peters, S. Vijayakumar, and S. Schaal, "Reinforcement learning for humanoid robotics," in *Proceedings of the third IEEE-RAS international conference on humanoid robots*, 2003, pp. 1–20.

[17] H. Sikchi, W. Zhou, and D. Held, "Learning off-policy with online planning," in *CoRL*, 2022.

[18] A. Kuleshov, A. Bernstein, and E. Burnaev, "Conformal prediction in manifold learning," in *Conformal and Probabilistic Prediction and Applications*. PMLR, 2018, pp. 234–253.

[19] S. Kiyani, G. Pappas, and H. Hassani, "Conformal prediction with learned features," *arXiv preprint arXiv:2404.17487*, 2024.

[20] A. Doula, T. Güdelhöfer, M. Mühlhäuser, and A. S. Guinea, "Conformal prediction for semantically-aware autonomous perception in urban environments," in *8th Annual Conference on Robot Learning*.

[21] J. Sun, Y. Jiang, J. Qiu, P. Nobel, M. J. Kochenderfer, and M. Schwager, "Conformal prediction for uncertainty-aware planning with diffusion dynamics model," *Advances in Neural Information Processing Systems*, vol. 36, pp. 80 324–80 337, 2023.

[22] A. Dixit, L. Lindemann, S. X. Wei, M. Cleaveland, G. J. Pappas, and J. W. Burdick, "Adaptive conformal prediction for motion planning among dynamic agents," in *Learning for Dynamics and Control Conference*. PMLR, 2023, pp. 300–314.

[23] L. Lindemann, M. Cleaveland, G. Shim, and G. J. Pappas, "Safe planning in dynamic environments using conformal prediction," *IEEE Robotics and Automation Letters*, vol. 8, no. 8, pp. 5116–5123, 2023.

[24] S. Yang, G. J. Pappas, R. Mangharam, and L. Lindemann, "Safe perception-based control under stochastic sensor uncertainty using conformal prediction," in *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE, 2023, pp. 6072–6078.

[25] J. Lekeufack, A. N. Angelopoulos, A. Bajcsy, M. I. Jordan, and J. Malik, "Conformal decision theory: Safe autonomous decisions from imperfect predictions," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 11 668–11 675.

[26] S. Sheng, P. Yu, D. Parker, M. Kwiatkowska, and L. Feng, "Safe pomdp online planning among dynamic agents via adaptive conformal prediction," *IEEE Robotics and Automation Letters*, 2024.

[27] H. Huang, S. Sharma, A. Loquercio, A. Angelopoulos, K. Goldberg, and J. Malik, "Conformal policy learning for sensorimotor control under distribution shifts," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 16 285–16 291.

[28] M. Zhao, R. Simmons, H. Admoni, and A. Bajcsy, "Conformalized teleoperation: Confidently mapping human inputs to high-dimensional robot actions," *arXiv preprint arXiv:2406.07767*, 2024.

[29] A. Kumar, J. Fu, M. Soh, G. Tucker, and S. Levine, "Stabilizing off-policy q-learning via bootstrapping error reduction," *NeurIPS*, 2019.

[30] A. Kumar, A. Zhou, G. Tucker, and S. Levine, "Conservative q-learning for offline reinforcement learning," *NeurIPS*, 2020.

[31] S. Fujimoto and S. S. Gu, "A minimalist approach to offline reinforcement learning," *NeurIPS*, vol. 34, pp. 20 132–20 145, 2021.

[32] S. Fujimoto, D. Meger, and D. Precup, "Off-policy deep reinforcement learning without exploration," in *ICML*, 2019.

[33] X. B. Peng, A. Kumar, G. Zhang, and S. Levine, "Advantage-weighted regression: Simple and scalable off-policy reinforcement learning," *arXiv preprint arXiv:1910.00177*, 2019.

[34] D. Garg, J. Hejna, M. Geist, and S. Ermon, "Extreme q-learning: Maxent rl without entropy," *arXiv preprint arXiv:2301.02328*, 2023.

[35] I. Kostrikov, A. Nair, and S. Levine, "Offline reinforcement learning with implicit q-learning," *arXiv preprint arXiv:2110.06169*, 2021.

[36] G. Williams, P. Drews, B. Goldfain, J. M. Rehg, and E. A. Theodorou, "Aggressive driving with model predictive path integral control," in *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2016, pp. 1433–1440.

[37] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International conference on machine learning*. PMLR, 2015, pp. 1889–1897.

[38] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *ICML*, 2018.

[39] Y. Lu, J. Fu, G. Tucker, X. Pan, E. Bronstein, R. Roelofs, B. Sapp, B. White, A. Faust, S. Whiteson *et al.*, "Imitation is not enough: Robustifying imitation with reinforcement learning for challenging driving scenarios," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 7553–7560.

[40] A. Nair, A. Gupta, M. Dalal, and S. Levine, "Awac: Accelerating online reinforcement learning with offline datasets," *arXiv preprint arXiv:2006.09359*, 2020.