Transformer-Based World Interaction Modeling for Humanoid Locomotion Control

Han Zheng[1,*], Yi Cheng[1,*], Hang Liu[2], Jiayi Li[1], Yizhe Li[1], Linqi Ye[3] and Houde Liu[1,†]

***Abstract*— Locomotion tasks for humanoid robots are challenging, especially in complex terrains. Understanding the physical processes of robot-environment interactions is key to achieving stable walking for humanoid robots. Since there is privileged information that the robot cannot directly access, the observation states are partially observable. Previous reinforcement learning(RL)-based methods either reconstruct environmental information from partial observations or reconstruct robotic dynamics information from partial observations, but they fail to fully model the physical processes of robot-environment interactions. In this work, we propose an end-to-end reinforcement learning control framework based on world physical interaction model for Humanoid robots. Our primary innovation is the introduction of a physical interaction world model to understand the dynamic interactions between the robot and the environment. Additionally, to address the temporal and dynamic nature of these interactions, we employ the hidden layers of Transformer-XL for implicit modeling. The proposed framework can showcase robust and flexible locomotion ability in complex environments such as slopes, stairs, and discontinuous surfaces. We validate the robustness of this method using the humanoid robot in simulations, and quantitatively compare our method against the baselines with better traversability and command-tracking.**

I. Introduction

Humanoid robots are expected to perform tasks related to human activities and collaborate with humans, which includes possessing motion capabilities comparable to humans and adapting their gaits to various terrain conditions. Although they exhibit superior mobility compared to wheeled robots in complex terrains, controlling them in scenarios with discontinuous contact and diverse motion skills remains challenging. Transitioning natural movements to humanoid robots still faces long-term technical challenges, including but not limited to the high degrees of freedom, underactuation, and complex non-linear dynamics of humanoid robots.

Traditional model-based control methods have significantly enhanced the locomotion capabilities of humanoid robots by using physical models to predict robot behavior [1]–[3]. However, these methods rely on accurate environmental dynamics modeling, which limits their application in complex terrains. Simplified dynamic models often lead to conservative movements, restricting the robot's potential. In contrast, RL-based methods [4]–[8] do not rely on detailed physical modeling and have shown greater flexibility and adaptability on legged robots. However, for humanoid robots, these methods can only handle relatively simple environments and have not yet fully addressed dynamic control issues in complex terrains.

Environmental information and robot motion information are essentially information from different domains, and result in understanding their interactions is challenging. Since actor networks can only obtain partial observations of the environment, they generally reconstruct partial observations into more complete environmental information by incorporating historical information or additional observational data. While these methods can reconstruct environment or robot dynamics information from partial observations, they fail to fully characterize the physical interaction processes between the robot and the environment. To address this issue, we introduce building world physical interaction model, which employ self-attention mechanisms to learn compact representations of historical observation inputs and implicitly infer latent interaction states by predicting future observation states.

Our input consists of temporally related historical sequence information, and we use the Transformer-XL [9], which allows the world model to directly access observations from previous time steps and learn long-term dependencies. The Transformer structure comprises multiple residual connections and self-attention mechanisms. The self-attention mechanism has unique advantages in modeling sequential information because it captures global information in the sequence without relying on fixed time windows.

We demonstrate the entire framework on the bipedal platform and validate our method. With our approach, the robot can traverse complex terrain in both simulation. Overall, our main contributions are:

- We propose a world physical interaction model, representing the first application of Transformer-XL based world model framework to humanoid robot tasks. By integrating it with the actor-critic method, we achieve enhanced RL exploration capabilities.
- Our approach incorporates time series information into the critic and leverages the world model for future predictions, significantly improving the critic network's ability to evaluate the robot's state and facilitating more globally informed decision-making.
- The simulation experiments demonstrate its superior traversability and command-tracking performance, fully showcasing the robustness of the approach.

* Equal Contributions

†Corresponding author: Houde Liu, liu.hd@sz.tsinghua.edu.cn.
This work was supported by the Natural Science Foundation of China under Grant 92248304 and the Shenzhen Science Fund for Distinguished Young Scholars under Grant RCJC20210706091946001.

[1]Tsinghua Shenzhen International Graduate School, Tsinghua University, 518055 Shenzhen, China.

[2]University of Michigan, Ann Arbor, MI 48109, USA.

[3]Shanghai University, 200444 Shanghai, China.

II. Related work

A. Blind Legged locomotion

For legged robot locomotion control, model-based methods are often difficult to generalize in an environment that is not modeled. Meanwhile, imitation learning [10]–[12] needs to rely on reference motion trajectories, but morphology and mass difference between human and robots result in scarce valid data. In contrast, RL can not only generalize to new environments, but also does not rely on reference trajectories. However, RL control also faces the challenge of Sim2Real Gap and limitation of perception, to solve this problem, there are a number of approaches [13]–[16]that utilize teacher-student strategy, with the teacher model receiving complete information. The output of the teacher model is then used to supervise the student model. In order to be able to better estimate privileged information that cannot be observed, some methods feed richer information such as gait [4]–[6]the controller, and some methods introduce state estimator modules [17]–[19], compensating for partial observability by expanding the state space. Our approach is also intended to enrich the observation space. However, by integrating a world model, we can better understand the deeper information embedded in the current observations—specifically, the interaction between the robot and its environment—through predictions of future observations.

B. World model for humanoid

The initial world model [20] is inspired by how humans process complex information to form an abstract representation of the world, understanding key entities and their interactions, and creating an internal representation of the world that allows predicting future events and making quick responses. For various problems that can be addressed using RL, the Dreamer series algorithms [21]–[23] have systematically explored the construction and learning of world models as well as the optimization of value and policy functions based on the actor-critic paradigm. Daydreamer [21] employs online learning, focusing on predicting future outcomes through experience with the world model and using these predictions to reduce the trial-and-error process in the actual environment, thereby improving training efficiency. The world denoising model [24] addresses the issue of discrepancies between simulation and real-world environments by utilizing the predictive capability of the world model for denoising. However, unlike the aforementioned methods, we innovatively apply the denoising model to abstract implicit features of the dynamical interaction between the robot and the environment, leveraging these features for decision-making and enabling robust locomotion in humanoid robots.

C. Transformers for humanoid

The Transformer [25] excels in handling long sequences and is compatible with various modalities and their combinations. It has achieved remarkable results in fields such as vision [26]–[28] and natural language processing [29]–[32]. In RL, decision-making methods such as Trajectory Transformer [33] and Decision Transformer [34] have been developed. For legged robot motion control tasks, [15] successfully deployed a control strategy to a quadrupedal robot by leveraging Decision Transformer and a two-stage knowledge distillation approach. [35] trained RL algorithms with a high-level vision controller to process visual and proprioceptive information and output target linear and angular velocities for driving lower-level controllers. [36], [37] used Transformers as feature extractors to achieve simple walking for humanoid robots. However, a common challenge in control tasks is that Transformers cannot capture the relationships between different segments, whereas our method using Transformer-XL establishes connections between different segments, avoiding information fragmentation.

Methods such as [38]–[42] combine Transformers with world models. Through this integration, We introduce a novel humanoid locomotion framework, in our method, transformers enable the world model to access past state information directly, rather than relying on compressed information, thus reducing the data compression process.

III. Method

A. Preliminary

1) Reinforcement learning task: In this paper, we formulate the problem of humanoid locomotion in complex terrain as a Partially Observable Markov Decision Process (POMDP) with discrete time steps $t \in \mathbb{N}$, defined as $\mathscr{M} = (S, A, O, T, Z, R, \gamma)$, where S, O, and A denote the state, observation, and action spaces, respectively.

The state transition probability $T(s', a, s)$ represents the probability of transitioning to a new state s' after executing action a in state s, defined as $T(s' \mid s, a) = P(s' \mid s, a)$. The observation probability $Z(o \mid s', a)$ represents the probability of observing o after executing action a and transitioning to a new state s'. The reward function $R(s, a)$ represents the expected reward obtained after executing action a in state s, and the discount factor $\gamma \in [0, 1)$ is used to weigh the relative importance of immediate rewards and future rewards. The ultimate goal is to find a policy π that maximizes the expected discounted reward:

$$\mathbf{J}(\pi) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\right]. \tag{1}$$

2) Task description: In our world physical interaction model, we decompose the locomotion task in complex environment into the following processes:

- **Dynamical Environment Understanding:** In complex environmental locomotion tasks, robot's understanding of its physical interactions with the environment determines its subsequent decisions. This process is highly dynamic and strongly temporally correlated, encompassing the relationship between environmental information and the robot's dynamic data, as well as the memory of these two types of information over historical time series and the robot's perception of environmental changes.

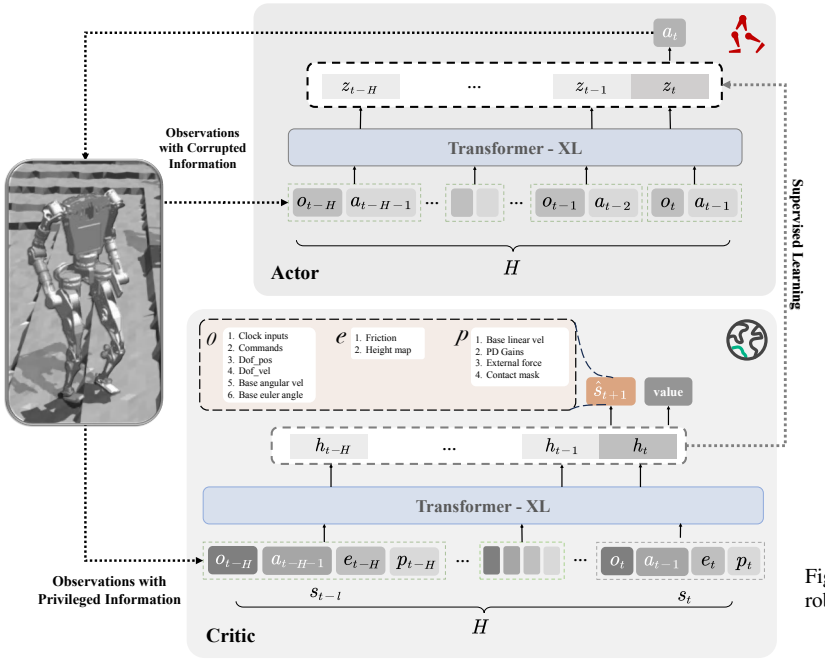


Fig. 1: Overview of world physical interaction model

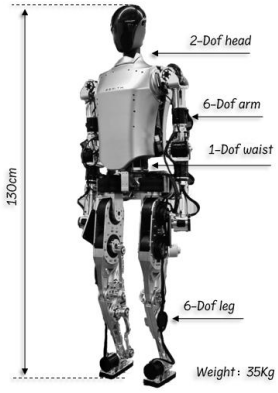


Fig. 2: Information about the humanoid robot.

- **Dynamic Prediction:** Dynamic prediction plays a pivotal role in enabling the robot to interact effectively with its environment. By leveraging interaction information, the robot can 'imagine' the complete states of the physical world and itself that would result from each possible action. This capability allows the robot to anticipate future dynamics and evaluate the potential value of its actions in a proactive manner. Such dynamic estimation not only enhances the robot's adaptability to diverse scenarios but also strengthens the generalization of its walking capabilities, particularly in complex and unpredictable environments.
- **Observation Space Expansion:** The robot can only access partially observable states of the environment. However, partial observations can only capture local information and fail to comprehensively represent the full complexity of the environment, making them insufficient to support the robot's decision-making requirements in complex environments. To learn comprehensive information, the policy network needs to expand the observation space based on historical observation sequences and dynamic predictions, ensuring that each extended state provides sufficient information to compensate for partial observability.

B. World Physical Interaction Model

1) Overview: Our proposed world physical interaction model method includes a dynamics model and a physical interaction regression model. We adopt an asymmetric actor-critic architecture, where the critic network combines with the dynamics model. The critic takes the historical observation state information s_t^H as input and compresses it into a hidden variable sequence h_t^H. The dynamics model predicts future state information $\hat{s}_{t+1}$, while the critic estimates the state value function. The actor network takes partial historical observation information $[o_t, a_{t-1}]^H$ as input and compresses it into a hidden variable sequence z_t^H. The actor hidden variables are supervised by the critic hidden variables to learn more interaction information. The actor network relies on

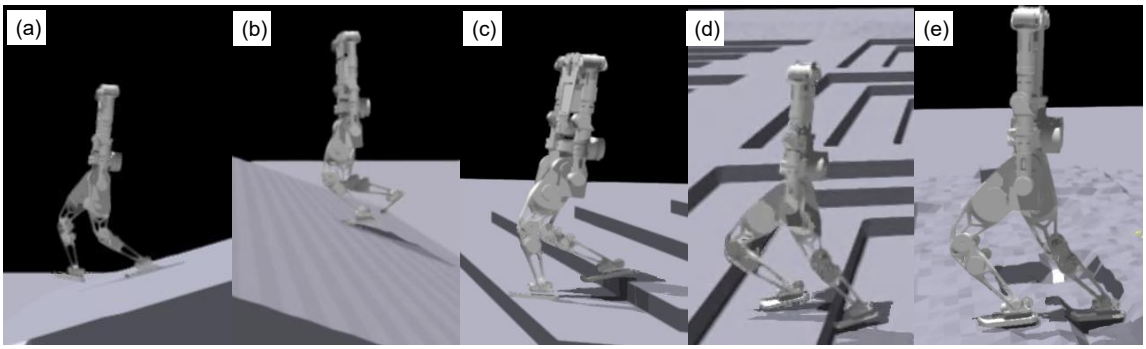


Fig. 3: The robot demonstrates robust locomotion across various terrains: a slope with an angle of 25° in (a) and (b), a staircase with a height of 10 cm in (c) and (d), and rough discontinuous surfaces in (e).

the value estimation provided by the critic network to make decisions. The Transformer-XL architecture allows the world model to directly access historical observation information rather than compressed information. Due to its recursive structure, each time step and previous hidden states together determine the current hidden state.

Observation Space: The observation space is composed of the following components:

- p_t: Represents partial environment and robot interaction information, including base linear velocity, PD gains, external force, and the contact state of the foot end.
- e_t: Represents environmental information, consisting of friction and the height map.
- o_t: Represents robot dynamics information, containing periodic signal input, desired velocity commands, joint position (q), joint velocity ($\dot{q}$), base angular velocity (ω_{xyz}), and base Euler angles in the coordinate system (θ_{xyz}).

Action Space: The dimension of the action spaces A equals the number of actuators. The movement of each actuator is formulated as the bias between the target joint position θ_{target} and the nominal joint position θ_0. The robot's target joint angle is defined as: $\theta_{\text{target}} = \theta_0 + ka_t$, where k is a scaling coefficient, k=0.25.

2) Dynamics Model: The dynamics model predict the next time state based on history observation state. Its backbone is an aggregation model f_ψ that compresses the observation state s_t^H into a hidden state sequence h_t^H. Using this hidden variable h_t, the dynamics estimation model predicts the next state $s_t\hat{+}1$. The dynamics model consists of these components:

$$\begin{aligned} \text{Aggregation Model}&: h_t^H = f_\psi(s_t^H) \\ \text{Dynamics Prediction Model}&: \hat{s}_{t+1} \sim p_\psi(\hat{s}_{t+1}|h_t) \end{aligned} \tag{2}$$

The aggregation model f_ψ is implemented as a causally masked Transformer-XL, while p_ψ employs a Multilayer Perceptron (MLP) to produce the output logits, these logits are subsequently converted into a one-hot representation that captures discrete categorical variables. Specifically, the output is discretized into 32 categorical variables, each capable of taking one of 32 possible categories. Transformer-XL introduces a recurrence mechanism that reuses the hidden states from the previous batch. This design overcomes the fixed-length limitation of traditional Transformer models, as highlighted in [43], allowing the model to process longer sequences efficiently. By integrating immediate dynamic changes from environmental interactions with long-term dependencies in time series, the model achieves enhanced predictive accuracy.

3) Physical Interaction Regression Model: We assume a). the critic can access the full observation of the environment, b). the hidden variable at time t has learned the historical observation information before time t. We believe that the latent variable h_t contains physical interaction information, and the regression model assists the actor network in learning this information. The regression model incorporates an aggregation model in (3), which encodes partial observation information $[o_t^H, a_{t-1}^H]$ into a hidden variable z_t^H.

$$\textit{Aggregation Model}: z_t^H = f_\psi(o_t^H, a_{t-1}^H) \tag{3}$$

f_ψ is also implemented as Transformer-XL. Specifically, as referenced in (5), physical interaction regression model employs a regression approach that utilizes the complete observation information provided by the critic network to guide the actor network in optimizing its latent variables. This process expands the observation space of the actor network, compensates for the limitations of partial observations, and enables the agent to better understand environmental dynamics and interaction relationships.

4) Policy learning: The actor network describes a Gaussian distribution based on the output mean and variance of the action, and then generates a specific action value by sampling from this distribution $a_t \sim \pi(a_t|o_t^H)$. The Critic network estimates the expected cumulative return R_t under the current policy at state s_t: $v_\psi(R_t \mid s_t)$. The key distinction from previous work lies in the introduction of time sequences and a world model for future prediction in our critic network not just actor network. This approach significantly enhances the critic's ability to evaluate the robot's state, thereby guiding decision-making with a more global perspective.

5) Loss Function: Our loss function includes the dynamics model loss, the reconstruction loss for hidden variable regression, and the policy optimization loss. In each iteration, we first update the dynamics model and the PPO module, followed by optimizing the regression module.

Dynamics Model Loss: Our goal is to ensure that the dynamics estimation model can accurately predict future observation state. Inspired by the balanced cross-entropy loss used in [40], we also calculate the entropy and cross-entropy. We use the cross-entropy L_{ent2} of the dynamics prediction model to prevent the encoder from deviating from the dynamics model. Entropy L_{ent1} regularizes the latent states and prevents them from collapsing into a one-hot distribution. The dynamics predictor L_{NLL} is optimized via negative log-likelihood, providing rich learning signals for the latent states.

$$\begin{aligned} L_{NLL} + L_{ent1} + L_{ent2} = \mathbb{E}\Bigg[\sum_{t=1}^{T} & \underbrace{-\ln p_\psi(\hat{s}_{t+1}|h_t)}_{\text{predictor}} \\ & + \underbrace{\alpha_1 H(h_t))}_{\text{entropy regularizer}} \\ & + \underbrace{\alpha_2 H(s_{t+1}, p_\psi(\hat{s}_{t+1}|h_t))}_{\text{consistency}} \Bigg] \end{aligned} \tag{4}$$

Hyperparameters α_1, α_2 are the relative weights of the terms.

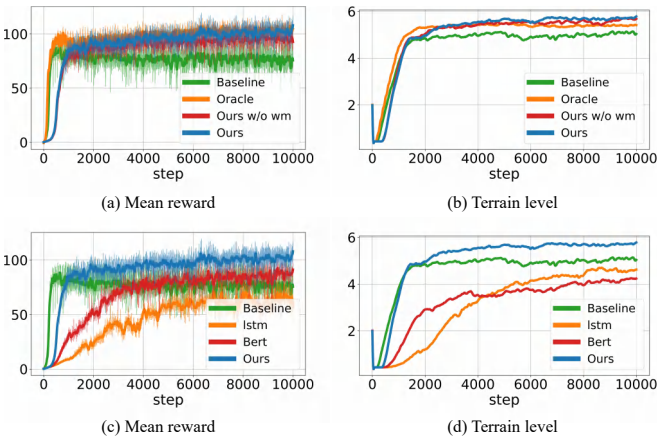
Reconstruction Loss: This loss corresponds to the regression model described in Section III-B.3, where the latent variable h_t generated by the critic network supervises the learning of the latent variable z_t produced by the actor

network. The mean squared error (MSE) loss we adopt for this purpose is as follows:

$$L_{reconstruct} = MSE(z_t, h_t) \tag{5}$$

Policy Optimization Loss: We use the Proximal Policy Optimization (PPO) algorithm to optimize the policy. The loss function primarily consists of a policy loss L^{clip} and a value function loss L_{value} The overall training loss is defined as

$$L = L^{clip} + L_{value} + L_{NLL} + L_{ent1} + L_{ent2} + L_{reconstruct} \tag{6}$$



(a) Mean reward (b) Terrain level

(c) Mean reward (d) Terrain level

Fig. 4: Comparison of different Method. (a) and (b) are compared with the baseline, oracle, and ablation experiments in terms of terrain levels and average rewards to demonstrate model performance, while (c) and (d) are compared with other methods to showcase the superiority of our model. We adopt curriculum learning [44] for training. Terrain level refers to the difficulty level of the terrain.

6) Training Details: We use the reward function as shown in Table I, where the task reward guides the robot to track the desired speed and complete motions on various terrains and alive reward mitigates the exploration burden in early period. Besides, we design comprehensive reward about feet [45], [4]to guide locomotion through tough terrain and prevent weird posture. Through extensive training trials, we optimize our reward weight settings to ensure that the robot moves in a relatively ideal manner. The domain randomizations and terrain setting details are in Table II and III.

IV. Experiments and Results

A. Experiment Setting

1) Benchmark Comparision: For a comparative evaluation, the experiments we performed are as follows:

- **Oracle:** Train the policy with a history of full privileged observations.
- **Baseline:** MLP network optimized using the PPO algorithm.
- **LSTM:** Adopt LSTM as network backbone [46]
- **Bert:** We implement the policy according to the Humanplus algorithm [36], Compared to Transformer-XL, the like-Bert structure lacks memory information and only focuses on the current time window.

TABLE I: Reward Function

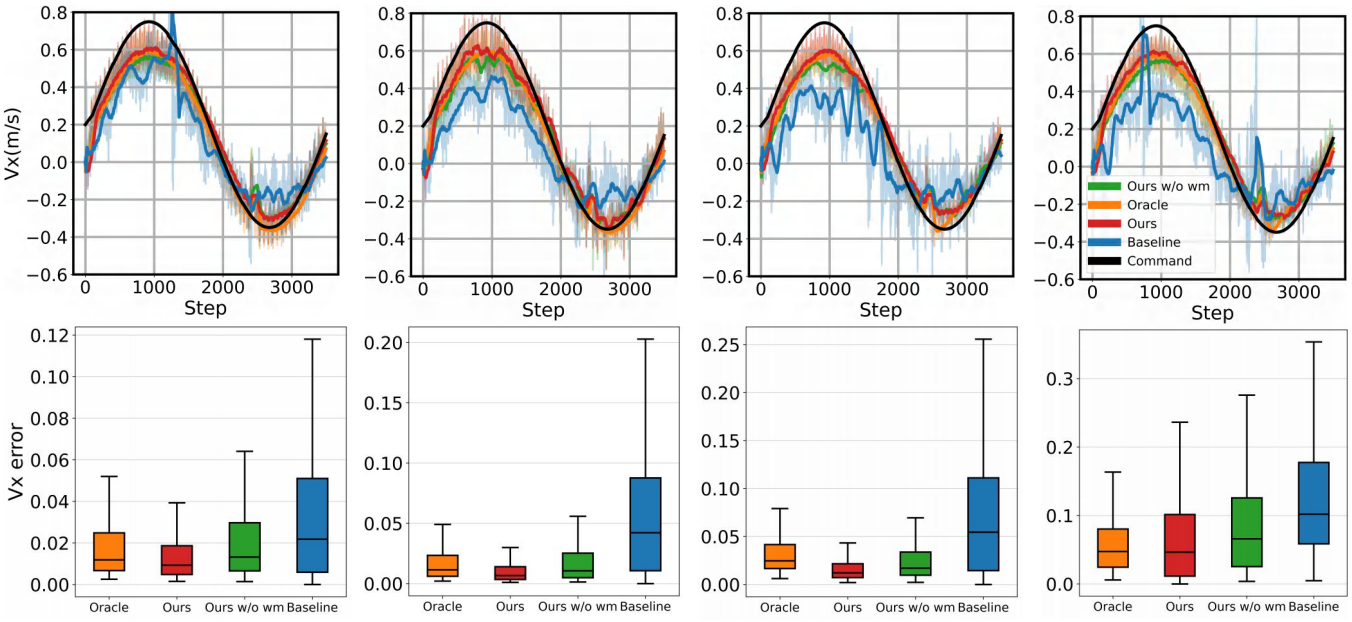
Term	Equation	Weight
Task Reward		
alive	1	0.5
xy velocity tracking	$\exp(-\lvert \mathbf{v}_{xy} - \mathbf{v}_{xy}^{cmd}\rvert^2 \cdot 5)$	1.5
yaw velocity tracking	$\exp(-(\omega_z - \omega_z^{cmd})^2 \cdot 5)$	1.0
Feet Guidance		
swing phase tracking (force)	$\sum_{foot}[1 - C_{foot}^{cmd}(\theta^{cmd}, t)] \exp(-\lvert \mathbf{f}^{foot}\rvert^2/100)$	5.0
stance phase tracking (velocity)	$\sum_{foot} C_{foot}^{cmd}(\theta^{cmd}, t) \exp(-\lvert \mathbf{v}_{xy}^{foot}\rvert^2/5)$	10.0
raibert heuristic tracking	$(\mathrm{p}_{xy,foot}^{f} - \mathbf{p}_{xy,foot}^{f,cmd}(s_y^{cmd}))^2$	-30.0
foot height tracking	$\sum_{foot}(h_{z,foot} - h_z^{f,cmd})^2 C_{foot}^{cmd}(\theta^{cmd}, t)$	-10.0
Regularization Reward		
body height	$\exp(-(h_z - h_Z^{cmd})^2 \cdot 1000)$	-0.2
z velocity	v_z^2	-0.02
foot slip	$\lvert \mathbf{v}_{xy}^{foot}\rvert^2$	-0.04
hip position	$\exp\left(-\sum_{i=1}^{2} q_{roll,yaw}^2 \cdot 100\right)$	0.4
feet orientation	$\exp\left(-\sum_{i=1}^{2} \lvert \theta_{roll,pitch}^{foot}\rvert \cdot 10\right)$	0.4
feet stumble	$\mathbb{1}\left(\max_i(\sqrt{F_{x_i}^2 + F_{y_i}^2} > 4\lvert F_{z_i}\rvert)\right)$	-1.0
orientation	$\exp(-\lvert g_{xy}\rvert^2 \cdot 10)$	1.5
thigh/calf collision	$1_{collision}$	-5.0
joint limit violation	$1_{q_i > q_{\max} \vee q_i < q_{\min}}$	-10.0
joint torques	$\lvert \tau\rvert^2$	-1e-5
joint velocities	$\lvert \dot{\mathbf{q}}\rvert^2$	-1e-3
joint accelerations	$\lvert \ddot{\mathbf{q}}\rvert^2$	-2.5e-7
action rate	$\lvert \mathbf{a}_t\rvert$	-5e-5
action smoothing	$\lvert \mathbf{a}_{t-1} - \mathbf{a}_t\rvert^2$	-0.01
action smoothing (2nd order)	$\lvert \mathbf{a}_{t-2} - 2\mathbf{a}_{t-1} + \mathbf{a}_t\rvert^2$	-0.01

TABLE II: Domain Randomizations and their Respective Range

Parameters	Range [Min, Max]	Unit
Ground Friction	[0.1, 1.5]	-
Ground Restitution	[0.0, 0.25]	-
Body Mass	[-2, 5]	Kg
Body Com	[-0.07, 0.1]	Kg
Link Mass	[0.8, 1.4] $\times$ nominal value	Kg
Joint K_p	[0.85, 1.15] $\times$ 20	-
Joint K_d	[0.85, 1.15] $\times$ 0.5	-
System Delay	[0, 40]	ms
External Force	interval = 5s $vel_{xy} = 0.4$	-

- **Ours w/o dynamics model:**The proposed method without dynamics estimation module.
- **Ours w/o regression model:** The proposed method without latent variable reconstruction.
- **Ours w/o world model:** The proposed method without dynamics estimation module and latent variable reconstruction.

2) Setups in Simulations: We conduct simulation experiments on the Isaac Gym platform, training 4096 agents in parallel using domain randomization. We test performance by comparing the convergence curves of rewards, the convergence curves of terrain levels, and the velocity tracking under various terrains. The training is conducted on an NVIDIA V100 GPU with 40 GB of memory. The detailed network hyperparameters are shown in Table IV.



(a) Discrete plane (b) Plane (c) Rough slope up (d) Stairs up

Fig. 5: Vehicle tracking comparision. We provide the robot with a sinusoidal velocity command and test the average velocity of 100 robots on different terrains. The Vx error is calculated using the following formula: $\text{Vx\_error} = \frac{1}{N}\sum_{i=1}^{N}\left(V_{x,\text{command}}(t) - V_{x,i}(t)\right)^2$.

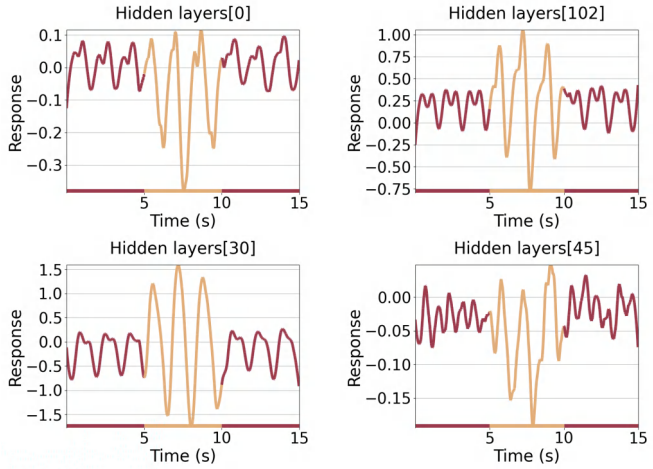


Fig. 6: Hidden layers visualization. The figure shows the changes in part of the hidden layer responses as the robot moves from flat ground to slope up and back to flat ground. The red line corresponds to the time when the robot is walking on flat ground, while the yellow line corresponds to the time when it is slope up.

TABLE III: Terrain Setting Range

Parameters	Range [Min, Max]	Proportion
Stair up	[5cm, 12cm]	0.5
Stair down	[5cm, 12cm]	0.5
Slope up	[0, 0.2]	2.5
Slope down	[0, 0.2]	1
Plane	-	0.5

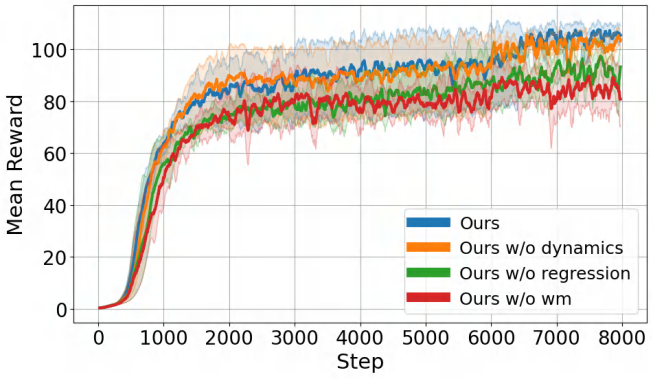
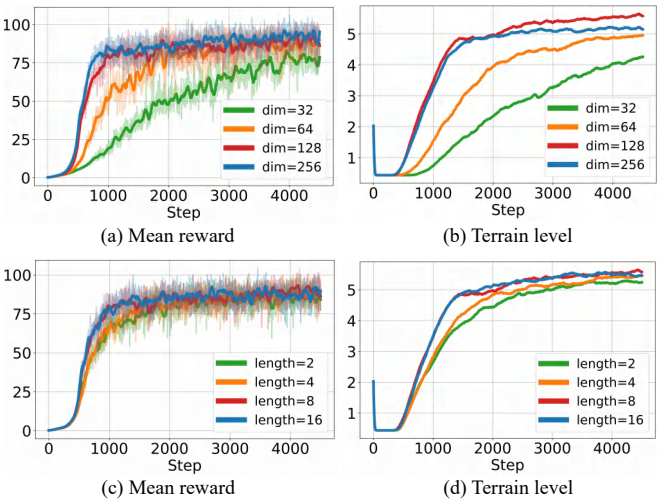


Fig. 7: Self-ablation experiments. We conduct three repeated experiments with different random seeds. The curves represent the average results, with shaded areas indicating the range between the minimum and maximum values.

B. Simulation Results

1) Terrain Passability Experiment: As shown in Fig.3, we test the upper limit and robustness of our method across various complex terrains. As shown in Fig. 4, our method significantly outperforms the baseline in handling complex terrains compared to the simple MLP structure. Additionally, our method surpasses even the "oracle" method, which has access to privileged information, in terms of the final terrain difficulty. This demonstrates that the transformer architecture effectively utilizes the robot's historical information to enhance decision-making. Our method also outperforms the ablated version, highlighting the importance of the world model in understanding dynamic interactions, allowing the robot to navigate complex terrains more efficiently and stably. The comparison with other methods further demonstrates that our approach is more robust and adaptable to different challenging terrains.

2) Command Tracking Experiment: We also quantitatively evaluated the ability of our method to track desired velocities in complex terrains. As shown in Fig. 5, (a), (b), (c), and (d) compare the velocity tracking performance



(a) Mean reward (b) Terrain level

(c) Mean reward (d) Terrain level

Fig. 8: The effect of the dimension of hidden layer and the time window length

TABLE IV: Hyperparameters of world physical interaction model

Parameter	Value
Number of Environments	4096
Context window	8
Memory window	8
Batch size	4096×24
Discount Factor	0.99
GAE discount factor	0.95
Entropy Coefficient	0.00001
PPO lr	0.0001
α_1	5.0
α_2	0.01
Transformer blocks	4
Embedding dimension	128
Multi-head attention heads	4
Reconstruction module lr	1×10^{-6}
Dynamic estimator module lr	1×10^{-6}

of different methods across various terrains. The top four plots show the actual velocity feedback curves as the robot tracks a continuously changing sine-wave desired velocity, while the bottom four plots present boxplots of the tracking errors in the x-direction for different methods. Our method demonstrates superior tracking performance across various terrains. In terms of both the upper bound of error and the median, our method significantly outperforms other methods. Even though the Oracle has access to foot elevation maps, our method outperforms Oracle on discrete, plane, and slope terrains. On stair terrains, which rely on foot elevation data, our method performs close to Oracle, indicating that the environmental estimation of our world model is already very close to the actual elevation map.

3) Latent Layer Analysis: As the robot transitions through a plane-slope-plane terrain environment, we visualized the outputs of 4 selected neurons from the 128-dimensional hidden layer. As shown in Fig. 6, the changes in hidden layer responses during terrain transitions highlight the robot's ability to adapt to varying terrains. These responses reflect the network's capability to recognize and respond to terrain changes, enabling real-time adjustments to the robot's control strategy.

C. Ablation experiment

As shown in Fig. 7, we compare our method with the ablated versions and found that the latent variable regression part and the future information prediction part influence each other. Having both components leads to better performance, which is understandable. The key to our approach lies in introducing time series through the critic and leveraging the world model for future predictions. This method enhances the evaluation capability of the critic network, guiding better decision-making abilities.

As shown in Fig. 8, we experiment with varying history length and hidden layer dimensions to verify whether our parameters achieve optimal locomotion performance and robustness. The time window determines the context range the model can observe when handling sequential tasks. A larger window helps capture long-range dependencies but increases computational costs. The model's performance is similar when the window length is 16 and 8, and significantly better than other window lengths. The hidden layer size determines the model's representation capacity, and increasing the number of hidden layers helps improve the network's fitting ability. The performance is similar when the number of hidden units is 256 and 128, with the model showing slightly better exploration ability in complex terrains when the hidden layer size is 128.

V. Conclusion

In this work, we propose world physical interaction model, a novel framework designed to address the challenges of humanoid robot locomotion in complex environments. Our framework integrates the world model concept into an asymmetric actor-critic structure, where the hidden layers of Transformer-XL implicitly model the dynamic interactions between the robot and its environment. This approach enhances decision-making by leveraging historical sequences and dynamic predictions to expand the observation space. We validate the effectiveness of our method through extensive simulation experiments, demonstrating its ability to achieve robust and adaptive locomotion across diverse and challenging terrains. The results highlight the advantages of incorporating world model-based implicit dynamics representation, allowing the robot to efficiently learn environment-aware control strategies. Future work will focus on improving model generalization to unseen terrains and exploring full-body coordination to enable more versatile and natural locomotion.

References

[1] P. M. Wensing and D. E. Orin, "Development of high-span running long jumps for humanoids," in *2014 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2014, pp. 222–227.

[2] M. Chignoli, D. Kim, E. Stanger-Jones, and S. Kim, "The mit humanoid robot: Design, motion planning, and control for acrobatic behaviors," in *2020 IEEE-RAS 20th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2021, pp. 1–8.

[3] M. S. Ahn, *Development and Real-Time Optimization-based Control of a Full-sized Humanoid for Dynamic Walking and Running*. University of California, Los Angeles, 2023.

[4] G. B. Margolis and P. Agrawal, "Walk these ways: Tuning robot control for generalization with multiplicity of behavior," in *Conference on Robot Learning*. PMLR, 2023, pp. 22–31.

[5] Z. Li, X. B. Peng, P. Abbeel, S. Levine, G. Berseth, and K. Sreenath, "Reinforcement learning for versatile, dynamic, and robust bipedal locomotion control," *arXiv preprint arXiv:2401.16889*, 2024.

[6] G. A. Castillo, B. Weng, W. Zhang, and A. Hereid, "Data-driven latent space representation for robust bipedal locomotion learning," *arXiv preprint arXiv:2309.15740*, 2023.

[7] J. Wu, G. Xin, C. Qi, and Y. Xue, "Learning robust and agile legged locomotion using adversarial motion priors," *IEEE Robotics and Automation Letters*, 2023.

[8] D. Kim, D. Carballo, J. Di Carlo, B. Katz, G. Bledt, B. Lim, and S. Kim, "Vision aided dynamic exploration of unstructured terrain with a small-scale quadruped robot," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 2464–2470.

[9] Z. Dai, "Transformer-xl: Attentive language models beyond a fixed-length context," *arXiv preprint arXiv:1901.02860*, 2019.

[10] A. Escontrela, X. B. Peng, W. Yu, T. Zhang, A. Iscen, K. Goldberg, and P. Abbeel, "Adversarial motion priors make good substitutes for complex reward functions," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 25–32.

[11] Z. Luo, J. Cao, K. Kitani, W. Xu, *et al.*, "Perpetual humanoid control for real-time simulated avatars," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 10 895–10 904.

[12] I. Radosavovic, B. Zhang, B. Shi, J. Rajasegaran, S. Kamat, T. Darrell, K. Sreenath, and J. Malik, "Humanoid locomotion as next token prediction," *arXiv preprint arXiv:2402.19469*, 2024.

[13] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning quadrupedal locomotion over challenging terrain," *Science robotics*, vol. 5, no. 47, p. eabc5986, 2020.

[14] A. Kumar, Z. Fu, D. Pathak, and J. Malik, "Rma: Rapid motor adaptation for legged robots," *arXiv preprint arXiv:2107.04034*, 2021.

[15] H. Lai, W. Zhang, X. He, C. Yu, Z. Tian, Y. Yu, and J. Wang, "Sim-to-real transfer for quadrupedal locomotion via terrain transformer," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 5141–5147.

[16] W. Wei, Z. Wang, A. Xie, J. Wu, R. Xiong, and Q. Zhu, "Learning gait-conditioned bipedal locomotion with motor adaptation," in *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)*. IEEE, 2023, pp. 1–7.

[17] G. Ji, J. Mun, H. Kim, and J. Hwangbo, "Concurrent training of a control policy and a state estimator for dynamic and robust legged locomotion," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4630–4637, 2022.

[18] I. M. A. Nahrendra, B. Yu, and H. Myung, "Dreamwaq: Learning robust quadrupedal locomotion with implicit terrain imagination via deep reinforcement learning," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 5078–5084.

[19] J. Long, Z. Wang, Q. Li, J. Gao, L. Cao, and J. Pang, "Hybrid internal model: A simple and efficient learner for agile legged locomotion," *arXiv preprint arXiv:2312.11460*, 2023.

[20] D. Ha and J. Schmidhuber, "World models," *arXiv preprint arXiv:1803.10122*, 2018.

[21] P. Wu, A. Escontrela, D. Hafner, P. Abbeel, and K. Goldberg, "Daydreamer: World models for physical robot learning," in *Conference on robot learning*. PMLR, 2023, pp. 2226–2240.

[22] D. Hafner, T. Lillicrap, M. Norouzi, and J. Ba, "Mastering atari with discrete world models," *arXiv preprint arXiv:2010.02193*, 2020.

[23] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap, "Mastering diverse domains through world models," *arXiv preprint arXiv:2301.04104*, 2023.

[24] X. Gu, Y.-J. Wang, X. Zhu, C. Shi, Y. Guo, Y. Liu, and J. Chen, "Advancing humanoid locomotion: mastering challenging terrains with denoising world model learning," *Robotics: Science and Systems*, 2024.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[27] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6836–6846.

[28] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[30] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, "Improving language understanding by generative pre-training," *None*, 2018.

[31] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[32] T. B. Brown, "Language models are few-shot learners," *arXiv preprint ArXiv:2005.14165*, 2020.

[33] F. Giuliari, I. Hasan, M. Cristani, and F. Galasso, "Transformer networks for trajectory forecasting," in *2020 25th international conference on pattern recognition (ICPR)*. IEEE, 2021, pp. 10 335–10 342.

[34] L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch, "Decision transformer: Reinforcement learning via sequence modeling," *Advances in neural information processing systems*, vol. 34, pp. 15 084–15 097, 2021.

[35] R. Yang, M. Zhang, N. Hansen, H. Xu, and X. Wang, "Learning vision-guided quadrupedal locomotion end-to-end with cross-modal transformers," *arXiv preprint arXiv:2107.03996*, 2021.

[36] Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn, "Humanplus: Humanoid shadowing and imitation from humans," *arXiv preprint arXiv:2406.10454*, 2024.

[37] I. Radosavovic, T. Xiao, B. Zhang, T. Darrell, J. Malik, and K. Sreenath, "Real-world humanoid locomotion with reinforcement learning," *Science Robotics*, vol. 9, no. 89, p. eadi9579, 2024.

[38] C. Chen, Y.-F. Wu, J. Yoon, and S. Ahn, "Transdreamer: Reinforcement learning with transformer world models," *arXiv preprint arXiv:2202.09481*, 2022.

[39] V. Micheli, E. Alonso, and F. Fleuret, "Transformers are sample-efficient world models," *arXiv preprint arXiv:2209.00588*, 2022.

[40] J. Robine, M. Höftmann, T. Uelwer, and S. Harmeling, "Transformer-based world models are happy with 100k interactions," *arXiv preprint arXiv:2303.07109*, 2023.

[41] W. Zhang, G. Wang, J. Sun, Y. Yuan, and G. Huang, "Storm: Efficient stochastic transformer based world models for reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[42] F. Deng, J. Park, and S. Ahn, "Facing off world model backbones: Rnns, transformers, and s4," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[43] C. Lu, R. Shi, Y. Liu, K. Hu, S. S. Du, and H. Xu, "Rethinking transformers in solving pomdps," *arXiv preprint arXiv:2405.17358*, 2024.

[44] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 41–48.

[45] J. Siekmann, Y. Godse, A. Fern, and J. Hurst, "Sim-to-real learning of all common bipedal gaits via periodic reward composition," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 7309–7315.

[46] J. Siekmann, S. Valluri, J. Dao, L. Bermillo, H. Duan, A. Fern, and J. Hurst, "Learning memory-based control for human-scale bipedal locomotion," 2020. [Online]. Available: https://arxiv.org/abs/2006.02402