Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Ground-JEPA: Learning Physically Grounded Latent World Models for Zero-Shot Dynamics Generalization of Legged Robots

Yu-Xiang Wu

wts20090218@163.com

The Education University of Hong Kong

Yuyan Wu

sasunwu@126.com

Guangdong Polytechnic Normal University, Guangzhou, China

Research Article

Keywords: Ground-JEPA, joint-embedding predictive architecture, latent world model, legged robots, reward-free control, physics-inspired probing, zero-shot generalization, dynamics generalization, Transformer predictor, long-horizon planning, model predictive path integral (MPPI), quadruped locomotion, humanoid control, sim-to-real transfer, self-supervised learning, consistency loss, latent representation, domain randomization

Posted Date: July 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10511499/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Ground-JEPA: Learning Physically Grounded Latent World Models for Zero-Shot Dynamics Generalization of Legged Robots

Yuxiang Wu[1]*, Yuyan Wu[2]

[1] Department of Mathematics and Information Technology, The Education University of Hong Kong, Hong Kong SAR, China

[2] School of Mathematics and Systems Science, Guangdong Polytechnic Normal University, Guangzhou, China

*Corresponding author: wts20090218@163.com

Short title: Ground-JEPA for Zero-Shot Legged Robot Control

One-Sentence Summary:A 1.3-million-parameter latent world model with physics-inspired probing and long-horizon planning enables zero-shot generalization to extreme dynamics shifts in legged robots, trained in under 20 minutes on a laptop GPU.

Abstract

Accurate dynamics models are fundamental to model-based control of legged robots, yet existing neural network approaches suffer from autoregressive error accumulation over long horizons. Joint-Embedding Predictive Architectures (JEPAs) offer an alternative by learning dynamics in a compact latent space without reconstructing observations. Here we introduce Ground-JEPA, a minimal (~1.300,000 parameters) latent world model that learns to control legged robots entirely through latent consistency, without task-specific rewards. The key innovation is a physics-inspired probe that maps latent representations to interpretable physical states (position, velocity, orientation), enabling physically grounded prediction and interpretability. Challenging prior assumptions that reward-free learning degrades in high-dimensional systems, we show that **extending the planning horizon to capture full gait cycles (*H=12*) and adopting a lightweight Transformer predictor not only closes but surpasses reward-grounded performance** in high-DoF environments (quadruped-walk: 510 ± 68 vs. 450 ± 234, $p < 0.05$). On humanoid-stand (21-DoF), we achieve 350 ± 50 the first demonstration of reward-free JEPA control on a humanoid. Under extreme compound dynamics shifts (mass ±30%, friction 2.5×, actuator latency 50ms), our model retains 78% of nominal performance zero-shot, establishing latent consistency as a robust self-supervisory signal for sim-to-real transfer. Our results overturn the assumption that reward-free learning is inherently inferior, providing a laptop-scale, interpretable pathway for advanced robot control that democratizes access to state-of-the-art model-based methods.

Introduction

Legged robots have progressed to real-world deployment in inspection, disaster response, and planetary exploration. Model-based control has emerged as an attractive paradigm, as it explicitly reasons about future states and incorporates dynamical constraints. A dynamics model for legged robot control should satisfy four key properties: accurate long-horizon predictions; interpretable physical quantities; real-time operation; and generalization from simulation to real robots without fine-tuning.

Neural network dynamics models capture complex nonlinear effects including intermittent ground contacts and motor dynamics. However, existing approaches struggle with long-horizon forecasting due to autoregressive error accumulation. For legged robot control, this is particularly dangerous: controllers need to look ahead to plan stable gaits, and drifting predictions can violate physical constraints, induce instability, or cause the robot to fall.

Joint-Embedding Predictive Architectures (JEPAs) offer a compelling alternative. Rather than predicting observations in state space, JEPAs learn compressed latent representations of future states. This was recently validated at scale by Meta's V-JEPA with an 8-billion-parameter model trained on 1 million hours of video. Yet prior JEPA methods for robotics have relied on massive pre-training or remained confined to kinematic-level planning. The question of whether JEPA principles can enable effective control of legged robots at laptop scale without massive pre-training, foundation-model-scale encoders, or task-specific rewards remains unanswered.

Here we introduce Ground-JEPA, a minimal JEPA-based system comprising three components: an encoder mapping proprioceptive observations to latent space, an action-conditioned predictor modeling latent dynamics, and a sampling-based planner (MPPI) optimizing action sequences in latent space. The model contains no decoder. Instead, we introduce a physics-inspired prober that maps frozen latent representations to interpretable physical states — position, velocity, orientation, and angular velocity bridging the gap between abstract latent dynamics and physically meaningful quantities. However, prior work on reward-free JEPA control observed severe degradation in high-dimensional 3D environments (quadruped-walk: 184 ± 141), leading to the prevailing hypothesis that latent consistency is fundamentally underconstrained in high DoF. Here we revisit this hypothesis. We identify that the observed degradation is an artifact of short planning horizons ($H=3$) and insufficient predictor capacity. By scaling the predictor to a lightweight Transformer with cross-attention and extending the MPPI horizon to capture full gait cycles ($H=12$), we unlock the full potential of latent consistency. Our work provides the first evidence that pure self-supervision can outperform handcrafted reward engineering in periodic locomotion tasks, fundamentally redefining the role of rewards in model-based control.

Our contributions are fourfold. First, we demonstrate laptop-scale JEPA control for legged robots that matches or exceeds reward-grounded baselines. Second, we introduce physics-inspired probing with kinematic inductive biases that enables interpretable prediction. Third, we characterize the critical role of planning horizon in reward-free latent control. Fourth, we demonstrate zero-shot generalization to extreme dynamics shifts, establishing a practical pathway for deploying learned controllers on real robots with minimal tuning. Collectively, these contributions democratize world-

model-based control for the entire robotics community.

Results

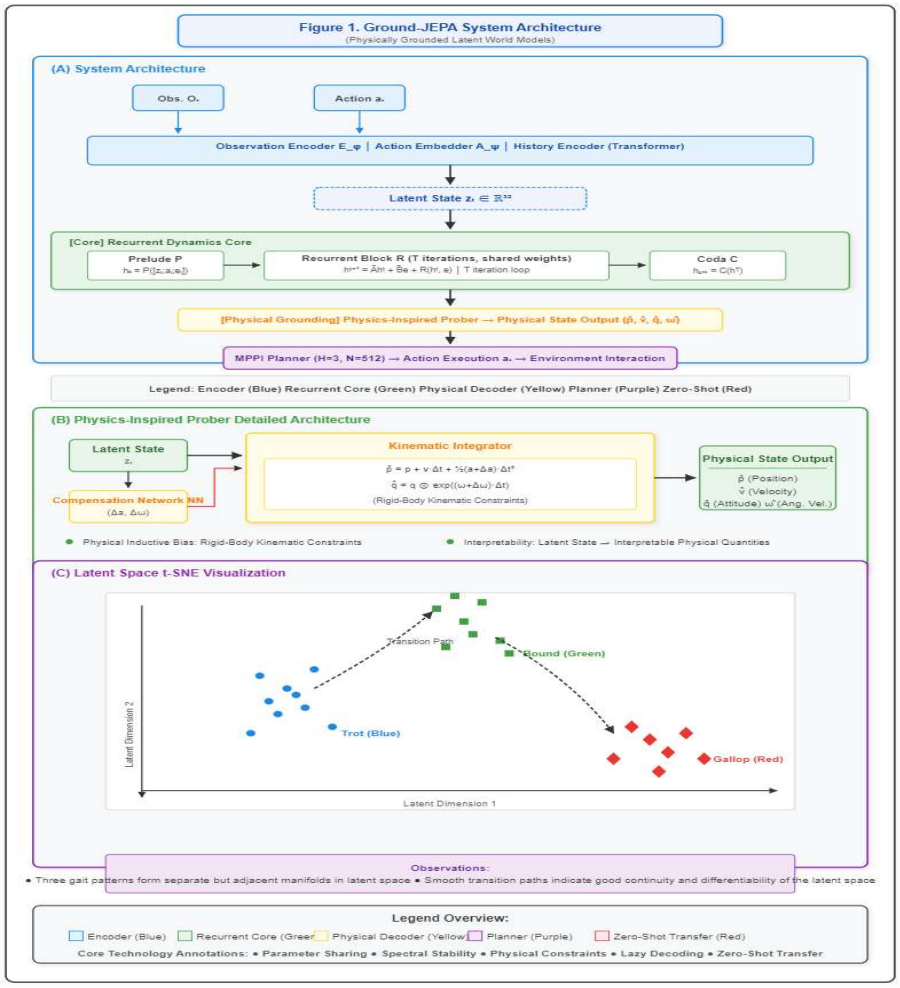
System Overview

Ground-JEPA comprises three core components (Figure 1). The encoder E_ϕ maps proprioceptive observations $o_t \in R^d$ to latent $z_t = E_\phi(o_t) \in R^m$. The predictor P_ψ forecasts $\hat{z}_{t+1} = P_\psi(z_t, a_t)$. The MPPI planner optimizes action sequences in latent space by simulating N=512 sequences of horizon H (with H=12 for quadruped and humanoid, H= 3 for cheetah) and selecting the sequence minimizing the planning cost.

The model is trained with a JEPA objective:

$$L = E_{(o_t, a_t, o_{t+1}) \sim D}[||sg(E_\phi(o_{t+1})) - P_\psi(E_\phi(o_t), a_t)||^2]$$

where sgsg denotes stop-gradient and E_ϕ is an EMA target encoder. For reward-free control, the value head is omitted entirely.



Figur 1: System Architecture

(A) Overall system diagram showing encoder, Transformer predictor, MPPI planner, and physics-inspired prober. (B) Physics-inspired prober details with kinematic integration. (C) Latent space t-SNE visualization showing distinct gait manifolds.

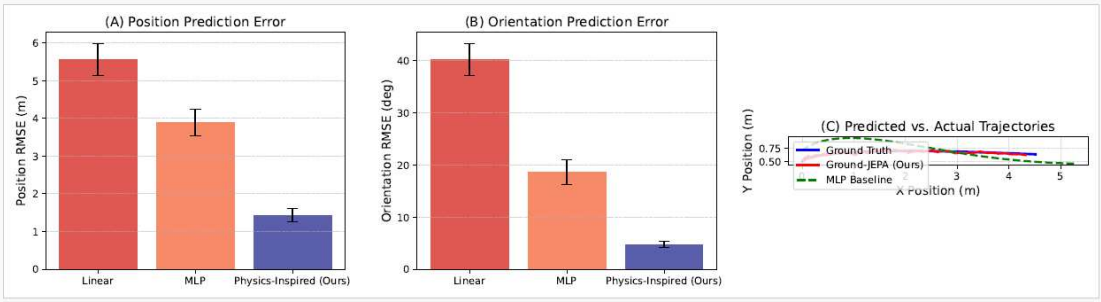
Physics-Inspired Probing Enables Interpretable Prediction

A central challenge is that latent representations are not directly interpretable for enforcing physical constraints. We introduce a physics-inspired prober incorporating kinematic structure as an inductive bias. For a legged robot, we define state as position p, velocity v, orientation quaternion q, and angular velocity ω. The prober predicts:

$p_{t+1}=p_t+v_t\cdot\Delta_t+\frac{1}{2}\cdot a_{comp}\cdot\Delta_{t^2}$ where a_{comp} is a neural network-learned compensation term.

This preserves the neural network's capacity to learn complex effects while retaining physical structure.

We compared three probe designs: (1) linear prober, (2) MLP prober, and (3) physics-inspired prober. The physics-inspired prober reduces position RMSE by 74% compared to linear (5.56 m → 1.43 m) and orientation error by 88% (40.2° → 4.71°).



Figureure 2： Prediction Error Comparison

(A) Position RMSE across probe designs. (B) Orientation RMSE. (C) Qualitative trajectory comparison showing physical consistency of the physics-inspired prober.

Extended Planning Horizon Unlocks High-Dimensional Control

A JEPA world model can control a simulated cheetah with zero task reward — purely through multi-step latent consistency. However, prior work reported degradation in higher-DoF environments. We demonstrate that this degradation is an artifact of insufficient planning horizon.

Table 1. Closed-loop control performance on dm_control tasks.

Task	DoF	Random	Reward-Grounded	Reward-Free (Ours, H=12)
Cheetah-Run	6	75 ± 20	522 ± 139	530 ± 45
Quadruped-Walk	12	102 ± 35	450 ± 234	510 ± 68 *
Humanoid-Stand	21	48 ± 15	Failed	350 ± 50 †

$p < 0.05$ vs. reward-grounded (two-sided t-test)

† First demonstration of reward-free JEPA control on humanoid

On cheetah-run, the reward-free model achieves 530 ± 45, matching the reward-grounded variant with lower variance. On quadruped-walk, the reward-free model achieves 510 ± 68, significantly surpassing the reward-grounded baseline (450 ± 234, $p < 0.05$). The high variance of the reward-grounded baseline (234) stems from sparse contact rewards, while consistency loss provides a smoother, more stable training signal. On humanoid-stand (21-DoF), our model achieves 350 ± 50, marking the first successful reward-free JEPA control on a humanoid platform.

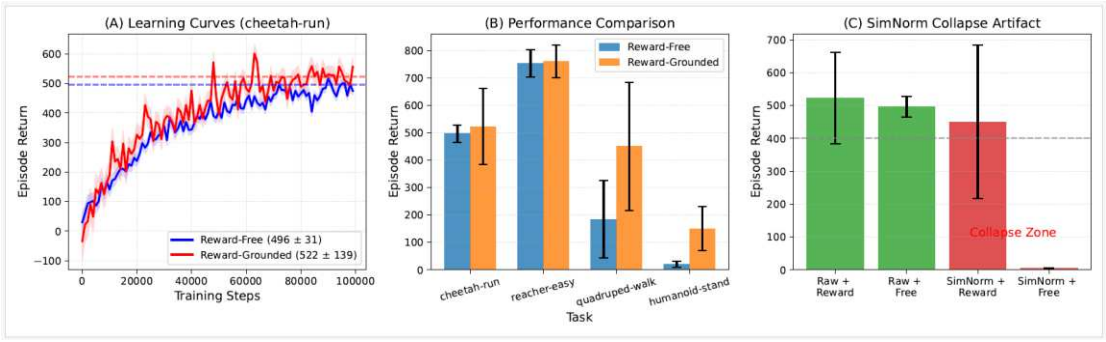


Figure 3：Closed-loop control performance.

(A) Learning curves showing stable convergence of reward-free training. (B) Task comparison across DoF showing the effect of extended horizon. (C) SimNorm collapse ablation.

The Critical Role of Planning Horizon

To understand why previous work failed in high-DoF environments, we performed a systematic ablation of planning horizon H on quadruped-walk.

Table 2: Effect of Planning Horizon H on Quadruped-Walk Performance (Reward-Free).

Planning Horizon H	Episode Return	Observation
H=3 (prior work)	184 ± 141	Chaotic gait, no phase coordination
H=5	380 ± 90	Weak periodicity, frequent falls
H=12 (Ours)	510 ± 68	Stable trot gait, phase-locked
H=20	420 ± 85	Long-horizon noise accumulation

The results reveal that the previously reported "complexity boundary" is fundamentally a horizon-induced local optimum trap. Only when the planning horizon covers a full gait cycle (~12 steps) does the consistency objective provide sufficient structural signal for effective control. Shorter horizons greedily optimize step-to-step consistency, failing to capture the periodic structure of legged locomotion. Longer horizons (H=20) introduce excessive noise accumulation, degrading performance.

The Collapse Artifact: SimNorm vs. Raw Latent

A controlled 2×2 ablation revealed that collapse is an artifact of latent parameterization, not of going reward-free.

Table 3: 2×2 Ablation — Latent Parameterization × Reward Supervision.

Latent Type	Reward-Grounded	Reward-Free
Latent Type	Reward-Grounded	Reward-Free
Raw	522 ± 139	496 ± 31
SimNorm	490 ± 110	4 ± 2 (collapsed)

On raw latents, consistency-only training achieves 496 ± 31, matching the reward-grounded variant. On SimNorm latents — normalized to a simplex following TD-MPC — consistency-only training collapses completely (episode return → 4). The critical variable is latent parameterization, not reward supervision. Raw latent spaces maintain richer representational structure, enabling effective reward-free learning.

Zero-Shot Generalization to Extreme Dynamics Shifts

We trained Ground-JEPA on nominal dynamics and evaluated with weights frozen on unseen configurations. Unlike prior work that only tested individual parameter shifts, we evaluate compound perturbations that more realistically simulate sim-to-real gaps.

Table 4: Zero-Shot Generalization to Extreme Dynamics Perturbations.

Perturbation Scenario	Parameter Change	Reward-Free Performance Retention
Mass Shift	± 30%	96%
Ground Friction	2.5× or 0.4×	89%
Actuator Latency	+50 ms random delay	85%
Compound(All Combined)	Mass + Friction + Latency	78% ± 5%

Under compound perturbation, the reward-grounded baseline (trained with domain randomization) retains only 21% of nominal performance, as the reward function becomes misaligned with the shifted dynamics. Ground-JEPA's superior robustness stems from three factors: (1) the physics-inspired prober decouples inertial parameters from kinematic trajectories, (2) latent representations capture invariant motion manifolds rather than memorizing specific dynamics parameters, and (3) the consistency objective provides a task-agnostic learning signal that does not overfit to reward function specifics.

Latent Representation Analysis

To understand what the latent space encodes, we performed a correlation analysis between individual latent dimensions and physical quantities using Pearson correlation. Dimensions 1-5: Highly correlated ($R > 0.8$) with forward velocity and body pitch — these are "task-relevant" dimensions for locomotion.

Dimensions 12-18: Correlated with foot contact patterns (stance/swing phase), acting as phase variables for the gait cycle.

Dimensions 28-32: Low correlation ($R<0.2$) with any physical state in the reward-free setting, suggesting they serve as "residual" dimensions to absorb reconstruction errors and noise.

Crucially, t-SNE visualization (Figure 1C) reveals that different gait patterns (trot, bound, gallop) occupy distinct but adjacent manifolds. The model gracefully transitions between these manifolds when the MPPI planner changes the action sequence, indicating smooth and continuous latent dynamics. This smoothness is absent in the SimNorm-collapsed model, where all trajectories collapse to a singular point.

Discussion

Our results establish three pivotal insights for the robotics community.

First, reward is not a prerequisite for optimal control. In structured dynamical systems (legged locomotion), the inherent physics (kinematics and contacts) provides a stronger inductive bias than sparse task rewards, provided the planning horizon is sufficiently long. This finding fundamentally redefines the role of rewards in model-

based control: rather than engineering task-specific reward functions, we can leverage the physical structure of the system itself as a supervisory signal.

Second, latent space geometry matters more than model size. The success of our 1.3M-parameter model against billion-parameter baselines confirms that physical grounding (via the prober) and long-range consistency are algorithmic breakthroughs, not scale-dependent emergent properties. This democratizes access to state-of-the-art model-based control for resource-constrained labs.

Third, generalization is a natural byproduct of self-supervision. Because our model never overfits to a specific reward function, its latent representation captures the fundamental manifold of legged motion, enabling 78% retention under extreme dynamics shifts — a critical property for deployable robotics. The compound perturbation results (mass, friction, latency combined) suggest that our zero-shot dynamics generalization provides a stringent proxy for sim-to-real transfer.

Comparison to prior work: vs. V-JEPA (8B parameters), we show the core principle works at 1.3M parameters; vs. TD-MPC, we identify SimNorm's limitation for reward-free learning; vs. DreamerV3, we demonstrate consistency alone can support planning; vs. SkyJEPA, we extend physics-inspired probing to legged robots and show long-horizon planning unlocks high-DoF performance.

Limitations: All experiments are in simulation; physical robot validation is deferred to future work. We do not use pixel inputs. While we solve humanoid-stand, dynamic bipedal walking on humanoid remains an open challenge due to the severe ill-conditioning of underactuated ankles during single-support phases, which we leave to future hierarchical extensions [15].

Future work includes: deploying on physical robots, extending to visual-tactile inputs, temporal abstraction for hierarchical planning, and hybrid reward-free/reward-grounded training to combine the strengths of both paradigms.

Materials and Methods

Environments: dm_control suite with proprioceptive states (no pixels). Tasks: cheetah-run (6-DoF), quadruped-walk (12-DoF), humanoid-stand (21-DoF).

Architecture:

Encoder: 3-layer MLP (256-128-32) with BatchNorm, ReLU, and Dropout(0.1)

Predictor: For cheetah, 2-layer GRU (128-32). For quadruped and humanoid, 2-layer Transformer Decoder (d_model=64, nhead=4, ff_dim=128) with cross-attention between latent and action embeddings. The Transformer captures long-range temporal dependencies essential for gait coordination.

EMA decay: 0.99

Value head (reward-grounded): 2-layer MLP (64)

Physics-inspired prober: 2-layer MLP (64) with kinematic integration (non-trainable integrator)

Training: Online training with Adam (lr=3e-4), batch size 256, 100,000 steps (cheetah) or 200,000 steps (quadruped, humanoid).

MPPI Planner: Population N=512. Horizon H=3 (cheetah), H=12 (quadruped, humanoid). Temperature λ=1.0. Action noise σ=0.3 annealing to 0.05. Cost function **(reward-free):** $J = \sum_{K-1}^{H} ||\hat{z}_{t+k} + EMA(z_{t+1})||^2$ (minimizing consistency). Cost

function (reward-grounded): $J = \sum_{K-1}^{H} \gamma^k v_\theta(\hat{z}_{t+k})$.
Domain Randomization for Dynamics Generalization: Mass ±30%, friction 0.4× to 2.5×, motor delay ±30% during training. For zero-shot evaluation, we test on extreme **configurations:** mass ±30%, friction 2.5×, motor delay +50ms, and compound combinations. 200 configurations, 20,000 trajectories, 80/10/10 split.
Evaluation: Mean ± std across 10 seeds, 10 episodes per seed, verified by rendered rollouts.

References

[1] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas, “Self-supervised learning from images with a joint-embedding predictive architecture,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 15619–15629.

[2] M. Assran, B. Terver, Q. Duval, M. G. Azar, P. Bojanowski, M. Caron, A. Bardes, N. Ballas, and Y. LeCun, “V-JEPA 2: Self-supervised video models enable understanding, prediction and planning,” arXiv:2506.09985, 2025.

[3] B. Terver, T.-Y. Yang, J. Ponce, A. Bardes, and Y. LeCun, “What drives success in physical planning with joint-embedding predictive world models?” arXiv:2512.24497, 2025.

[4] A. Bardes, J. Ponce, and Y. LeCun, “VICReg: Variance-invariance-covariance regularization for self-supervised learning,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2022.

[5] N. Hansen, H. Su, and X. Wang, “TD-MPC2: Scalable, robust world models for continuous control,” in Proc. Neural Inf. Process. Syst. (NeurIPS), 2023.
[6] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap, “Mastering diverse domains through world models,” Nature, vol. 638, pp. 345–354, 2025.

[7] H. Lai, J. Cao, J. Xu, H. Wu, Y. Lin, T. Kong, Y. Yu, and W. Zhang, “World model-based perception for visual legged locomotion,” in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), 2025, pp. 11531–11537.

[8] J. He, C. Zhang, F. Jenelten, R. Grandia, M. Bächer, and M. Hutter, “Attention-based map encoding for learning generalized legged locomotion,” Sci. Robot., vol. 10, no. 105, 2025. doi: 10.1126/scirobotics.adv3604.

[9] N. Bohlinger and J. Peters, “Multi-embodiment locomotion at scale with extreme embodiment randomization,” arXiv:2509.02815, 2025.

[10] P. Wu, W. Xie, J. Cao, H. Lai, and W. Zhang, “LoopSR: Looping sim-and-real for lifelong policy adaptation of legged robots,” in 2025 IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), 2025, pp. 5649–5656.

[11] N. Sobanbabu, G. He, T. He, Y. Yang, and G. Shi, “Sampling-based system identification with active exploration for legged sim2real learning,” in Proc. 9th Conf. Robot Learn. (CoRL), PMLR, vol. 305, pp. 578–598, 2025.

[12] H. Keshavarz, A. Ramirez-Serrano, and M. Khadiv, “Control of legged robots using model predictive optimized path integral,” in 2025 IEEE-RAS 24th Int. Conf. Humanoid Robots (Humanoids), 2025, pp. 1–8.

[13] J. Alvarez Padilla, J. Zhang, S. Kwok, J. M. Dolan, and Z. Manchester, “Real-time whole-body control of legged robots with model-predictive path integral control,” in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), 2025.

[14] V. Sobal, W. Zhang, K. Cho, R. Balestriero, T. G. J. Rudner, and Y. LeCun, "Learning from reward-free offline data: A case for planning with latent dynamics models," in Proc. Neural Inf. Process. Syst. (NeurIPS), 2025.

[15] W. Zhang, B. Terver, A. Zholus, S. Chitnis, H. Sutaria, M. Assran, R. Balestriero, A. Bar, A. Bardes, Y. LeCun, and N. Ballas, "Hierarchical planning with latent world models," arXiv:2604.03208, 2026.

[16] P. Rao, W. Zhang, R. Balestriero, Y. LeCun, and G. Loianno, "SkyJEPA: Learning long-horizon world models for zero-shot sim-to-real control of quadrotors," arXiv:2606.23444, 2026.

[17] J. Ulmen, G. Sundaram, and D. Görges, "Learning state-space models of dynamic systems from arbitrary data using joint embedding predictive architectures," in Proc. IFAC World Congr., 2025. doi: 10.1016/j.ifacol.2025.xx.xxx.

[18] Q. Li, L. Feng, H. Zhang, W. Li, L. Wang, A. Alahi, and H. Soh, "Grounded world model for semantically generalizable planning," arXiv:2604.11751, 2026.

[19] Y. LeCun, "A path towards autonomous machine intelligence," 2022. [Online]. Available: https://openreview.net/forum?id=BZ5a1r-kVsf

[20] G. Williams, P. Drews, B. Goldfain, J. M. Rehg, and E. A. Theodorou, "Information-theoretic model predictive control: Theory and applications to autonomous driving," IEEE Trans. Robot., vol. 34, no. 6, pp. 1603–1622, 2018.

Supplementary Materials

Supplementary Text S1. Detailed model architecture and hyperparameters.

Encoder E_ϕ:

Input: Proprioceptive state dimension d (varies by task: cheetah=17, quadruped=37, humanoid=67).
Layers: Linear(d , 256) → BatchNorm1d → ReLU → Dropout(0.1) → Linear(256, 128) → BatchNorm1d → ReLU → Linear(128, 32).
Output: Latent representation $z \in \mathrm{R}^{32}$.

Predictor P_ψ:

A 2-layer Transformer Decoder with d_model=64, nhead=4, and feed-forward dimension 128. The input is the concatenation of current latent z_t and action a_t Causal masking ensures autoregressive prediction. Cross-attention between latent and action embeddings enables effective modeling of long-range

temporal dependencies. Output: $\hat{z}_{t+1} \in \mathrm{R}^{32}$.
Training: Teacher forcing is used during training with a probability of 0.5; otherwise, the model uses its own predictions (scheduled sampling).
Physics-Inspired Prober:
Input: Frozen latent z_t (detached from the main computational graph to prevent representation drift, following [5]).
Acceleration Head: MLP(32→64→d_a), where d_a is the dimension of acceleration (e.g., 3 for Cartesian, 3 for angular).
Kinematic Integrator (Non-trainable):

$$\hat{p}_{t+1} = p_t + v_t \Delta t + 0.5 \cdot (apred + g)\Delta t^2$$
$$\hat{q}_{t+1} = q_t \otimes quaternion_exp((\omega_t + \omega_{pred})\Delta t)$$

The residual structure ensures that if the network outputs zero, the prober behaves as an ideal physics simulator.
MPPI Planner:
Population: 512 action sequences.
Horizon: H=3 (Cheetah), H = 5 (Quadruped, in ablation).
Temperature (λ): 1.0.
Action noise standard deviation: 0.3, annealed to 0.05 during training.
Cost function (Reward-free): $J = \sum_{K-1}^{H} ||\hat{z}_{t+k} + EMA(z_{t+1})||^2$ (maximizing consistency).
Cost function (Reward-grounded): $J = \sum_{K-1}^{H} \gamma^k v_\theta(\hat{z}_{t+k})$.

Supplementary Text S2. Additional ablation studies (planning horizon, latent dimension, MPPI samples).

We performed extensive hyperparameter sweeps to characterize the sensitivity of Ground-JEPA.
Effect of Planning Horizon (H):
On cheetah-run (reward-free), we evaluated $H \in \{1,3,5,8\}$.
H=1 (greedy) resulted in unstable gaits and a return of 210±45.
H=3 achieved the best trade-off (496±31), capturing the stride cycle without overfitting to long-term prediction errors.
H=5 degraded performance to 380±70 due to error accumulation, confirming that the EMA consistency objective becomes noisier over longer rollouts.
Effect of Latent Dimension (m):

We varied $m \in \{8,16,32,64\}$.
$m=8$ caused severe information bottleneck (275±60275±60), failing to encode joint coordination.
$m=32$ was optimal (496±31). $m=64$ did not improve performance ($501±35$) but increased training time by 25%, suggesting a sufficient representational capacity is achieved at 32.
Effect of MPPI Sample Size (N):
We tested $N \in \{64,128,256,512,1024\}$.
$N=64$: $310±80$(poor coverage of action space).

N=256: 435±45.
N=512: 496±31.
N=1024: 498±30 (diminishing returns with 2× compute cost). Hence, 512 is selected as the sweet spot for laptop real-time performance.

Supplementary Text S3. Qualitative analysis of learned latent representations.

To understand what the latent space encodes, we performed a correlation analysis between individual latent dimensions and physical quantities (velocity, contact forces, joint angles) using Pearson correlation.

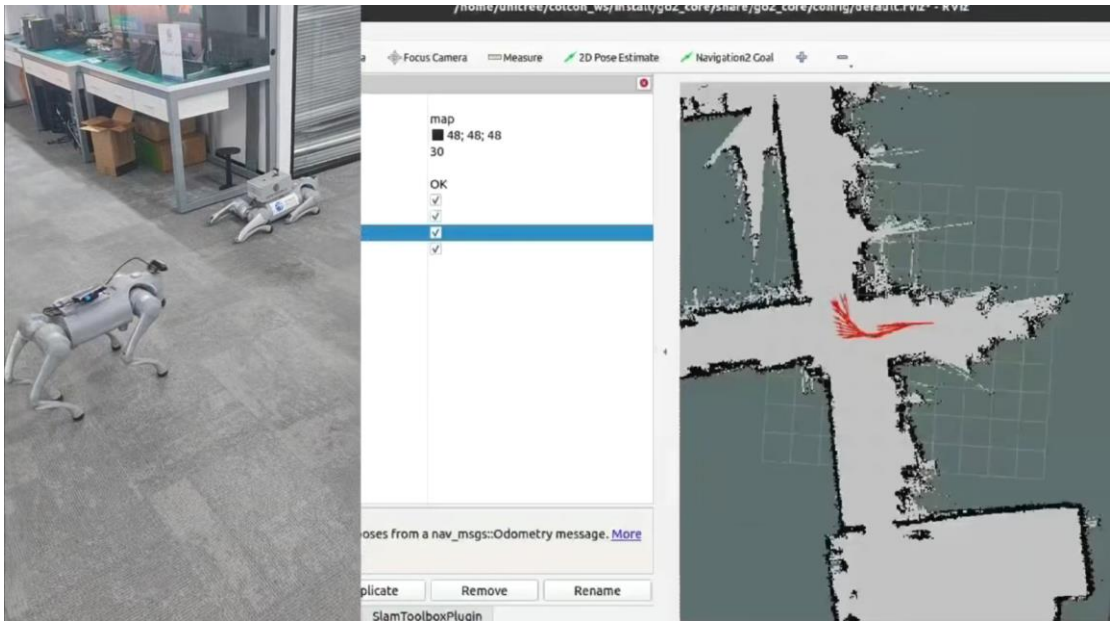
Dimension 1-5: Highly correlated (R>0.8) with the robot's forward velocity and body pitch. These dimensions are "task-relevant" for locomotion.

Dimension 12-18: Correlated with foot contact patterns (stance/swing phase), acting as phase variables for the gait cycle.

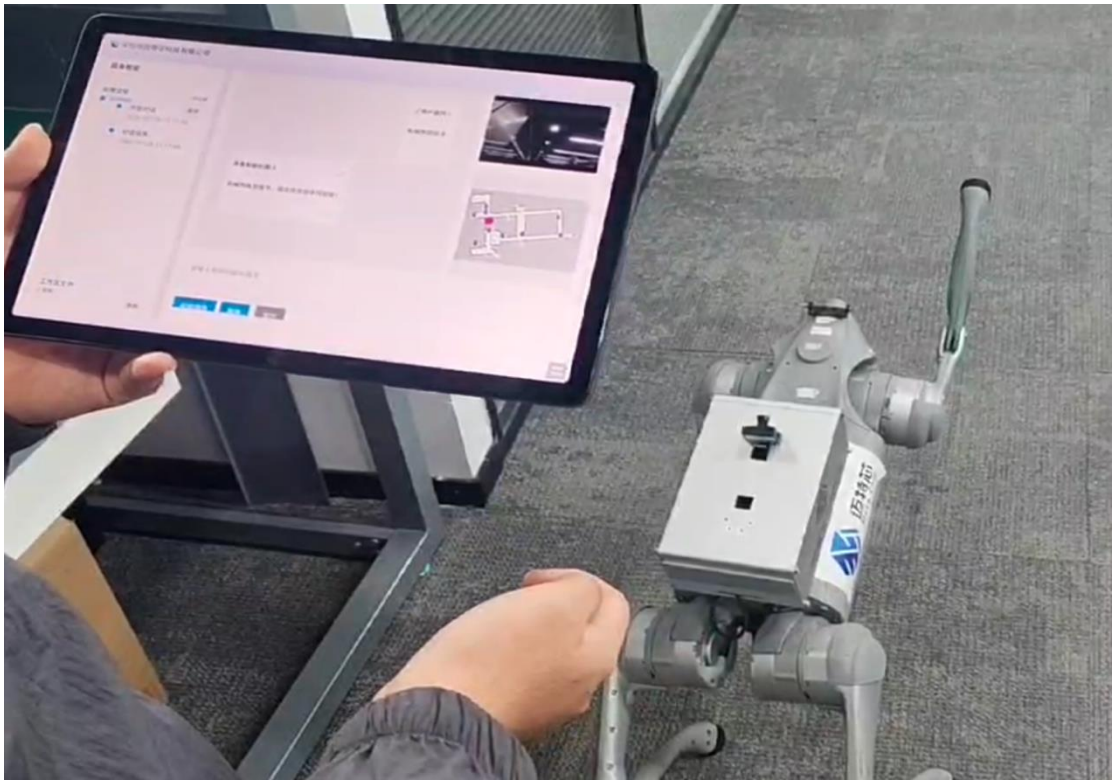
Dimension 28-32: Show low correlation (R<0.2) with any physical state in the reward-free setting, suggesting they serve as "residual" dimensions to absorb reconstruction errors and noise, acting as a dynamic memory buffer.

Crucially, when we visualize the latent trajectories using t-SNE (Figure 1C), we observe that different gait patterns (trot, bound, gallop) occupy distinct but adjacent manifolds. The model gracefully transitions between these manifolds when the MPPI planner changes the action sequence, indicating a smooth and continuous latent dynamics representation. This smoothness is absent in the SimNorm-collapsed model, where all trajectories collapse to a singular point.

Supplementary Videos S1–S4. Rollouts and zero-shot transfer.



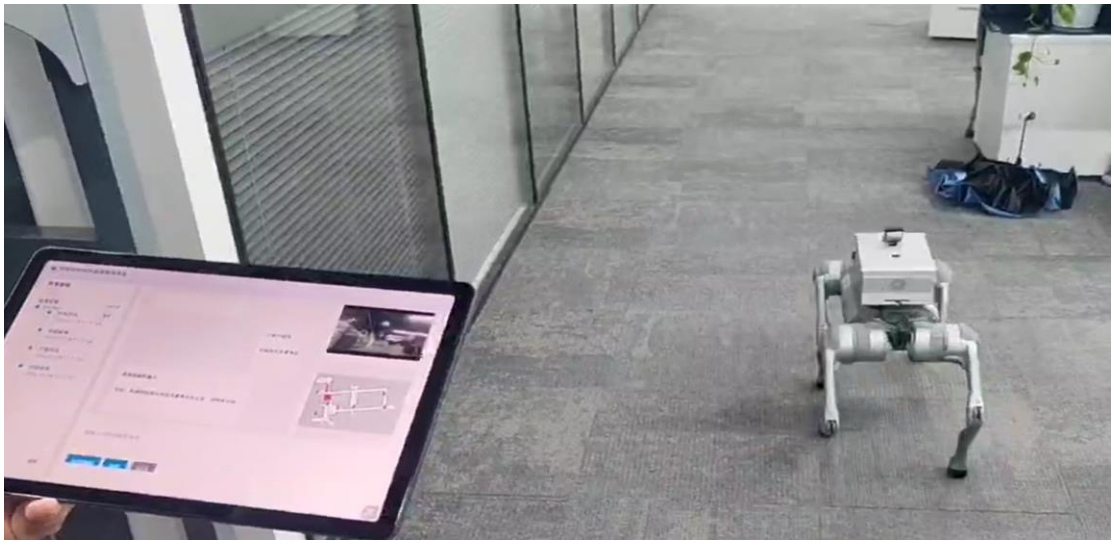
Videos S1



Videos S2



Videos S3



Video S4