Learning to Walk in Costume: Adversarial Motion Priors for Aesthetically Constrained Humanoids

Arturo Flores Alvarez[1], Fatemeh Zargarbashi[2], Havel Liu[1], Shiqi Wang[1],
Liam Edwards[1], Jessica Anz[1], Alex Xu[1], Fan Shi[3], Stelian Coros[2], Dennis W. Hong[1]

***Abstract*— We present a Reinforcement Learning (RL)-based locomotion system for Cosmo, a custom-built humanoid robot designed for entertainment applications. Unlike traditional humanoids, entertainment robots present unique challenges due to aesthetic-driven design choices. Cosmo embodies these with a disproportionately large head (16% of total mass), limited sensing, and protective shells that considerably restrict movement. To address these challenges, we apply Adversarial Motion Priors (AMP) to enable the robot to learn natural-looking movements while maintaining physical stability. We develop tailored domain randomization techniques and specialized reward structures to ensure safe sim-to-real, protecting valuable hardware components during deployment. Our experiments demonstrate that AMP generates stable standing and walking behaviors despite Cosmo's extreme mass distribution and movement constraints. These results establish a promising direction for robots that balance aesthetic appeal with functional performance, suggesting that learning-based methods can effectively adapt to aesthetic-driven design constraints.**

I. Introduction

Humanoid robots play an increasingly vital role in human environments thanks to their intuitive interaction capabilities and adaptability. Their anthropomorphic design equips them to navigate and perform tasks within spaces designed for humans, ranging from simple locomotion to intricate manipulation tasks [1]. Particularly, in the entertainment sector, humanoid robots enrich storytelling and audience engagement through lifelike movements, significantly enhancing performances across films, theme parks, and live shows. [2], [3].

Entertainment humanoids face unique challenges: disproportionate body parts shifting center of mass, restricted sensing from aesthetic shells, and limited joint mobility from protective coverings—all prioritizing visual appeal over stability.

Our robot, Cosmo, encompasses these challenges: it lacks an onboard vision system with an elevated center of mass, making it an ideal case study for testing the limits of conventional locomotion methods and exploring the capabilities of learning-based solutions.

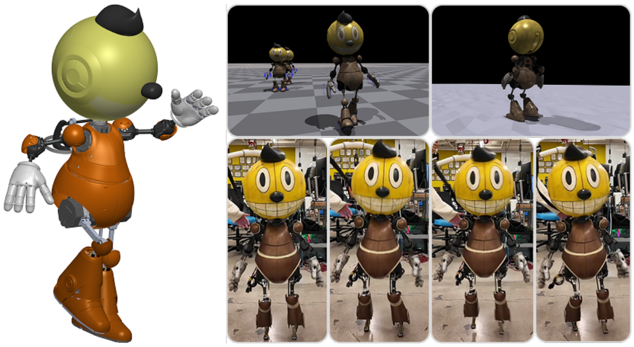


Fig. 1: Cosmo: an entertainment humanoid robot with covers designed for a blockbuster movie. (**Left**): CAD Design. (**Top**): Using Isaac Gym's massively parallelized environments to train with different styles and terrain. (**Bottom**): Sim-to-Real demonstration of natural walking (see supplementary video).

In this work, we demonstrate how modern learning-based methods, particularly Adversarial Motion Priors (AMP) [4], can overcome these difficulties. AMP is a method that blends the realism of imitation learning with the flexibility of reinforcement learning, enabling robots to develop physically plausible movements while retaining the signature style of human movements. We begin by drawing on human motion capture data from established datasets, then retargeting it onto our custom robot character. This step ensures the resulting movements preserve the essence of human behavior while adapting to the unique physical constraints of our platform.

By combining AMP with comprehensive sim-to-real strategies—including domain randomization and meticulous hardware parameter tuning—we achieve real-world locomotion on Cosmo, addressing challengs in balancing, walking, and safe deployment. Our key contributions are:

1) A learning-based locomotion control system for a humanoid with an extreme mass distribution and shifted center of mass—a configuration rarely addressed in humanoid control literature.
2) A sim-to-real transfer pipeline tailored for robots with aesthetic shell constraints, enabling safe transfer from simulation to physical hardware while preserving both performance and component safety.
3) A demonstration that AMP-guided reinforcement learning can generate natural, stable walking behaviors even on platforms with significant mechanical limita-

[1] Robotics and Mechanisms Laboratory (RoMeLa), Dept. of Mechanical and Aerospace Engineering, UCLA, CA, USA.

[2] Computational Robotics Lab (CRL), Dept. of Computer Science, ETH Zürich, Switzerland.

[3] Human-Centered Robotics Lab, Dept. of Electrical and Computer Engineering, National University of Singapore, Singapore.

The authors thank Min Sung Ahn for his expertise in sim-to-real transfer and best practices on hardware, Norman Zhu for sharing model-based walking experimental data on Cosmo, Conrad Ku and Premkumar Sivakumar for their support during testing phases, and Yusuke Tanaka and Alvin Zhu for their insights on improved training methodologies.

tions and entertainment-focused design constraints.

This work addresses an emerging challenge in entertainment robotics: achieving stable and expressive locomotion in designs where aesthetics dominate over functionality. By showcasing how learning-based methods can overcome these unique constraints, we highlight a promising path forward for robots that must balance visual appeal with real-world performance. Our results suggest that learning-based methods, especially when guided by motion priors, offer a powerful tool for enabling dynamic motion in robots destined for rich, interactive entertainment environments.

II. Related Work

Model-based approaches have long been the foundation of humanoid locomotion [5], [6]. These controllers typically require accurate models and extensive manual tuning, and they are often brittle when conditions deviate from those assumed in the design. In contrast, reinforcement learning (RL) has emerged as a powerful alternative for humanoid control [7]–[9]. Pure RL approaches, such as [10]–[12], have demonstrated impressive dynamic capabilities, enabling agile behaviors like running and parkour. These methods often assume well-structured morphologies and full-state sensing.

Despite their success, purely RL-based methods typically require extensive reward shaping and often produce unnatural, non-human-like movements due to the absence of imitation signals. Imitation learning offers a way to imbue humanoid controllers with human-like movement qualities by learning from demonstration data (such as motion capture of human locomotion). A significant practical advantage of this approach is that it establishes clear visual benchmarks for expected behavior—when the robot deviates from reference motions, these discrepancies can be quickly identified and debugged through visual inspection. In particular, Peng et al. [4] introduced Adversarial Motion Priors (AMP) to incorporate motion data into RL training loop via adversarial training, producing natural walking and running motions. [13] further demonstrated that AMP can replace complex reward engineering by learning natural motion behaviors through adversarial training alone.

Our work builds upon these methods but differs in two key ways: (1) our hardware is significantly more difficult to control due to its top-heavy design, and (2) our system lacks visual input, forcing the policy to rely entirely on proprioception. While imitation learning and domain randomization are well-studied in sim-to-real transfer, our results demonstrate their efficacy under harder conditions.

III. Methods

A. Motion Retargeting

The increasing availability of human motion capture data has become a natural and effective choice for generating reference motions for lifelike robots. However, a significant challenge arises due to differences in morphology. For example, Cosmo's legs rotate at two points in the hip, whereas a human just has a ball joint at the hip. Therefore, an intermediate process is necessary to adapt human motion data to the kinematic constraints of a specific robotic platform. This process is commonly known as *retargeting*.

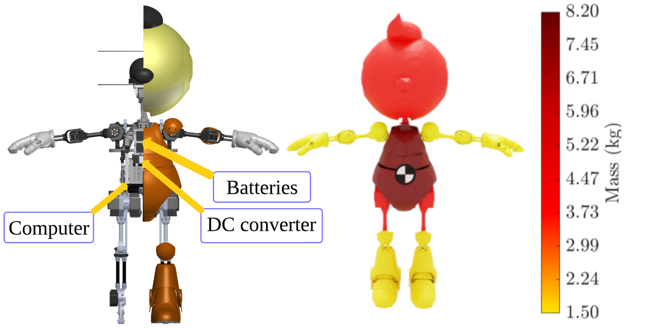


Fig. 2: Cosmo visualization: (left) arm range of motion and internal vs. exterior housing comparison; (right) mass distribution analysis highlighting the disproportionate head mass.

We utilize the Rokoko plugin [14] in Blender [15] for retargeting CMU Mocap Dataset [16], [17]. A custom animation rig matching the Cosmo robot's proportions is built in Blender. Human motion data is approximated by averaging limb movements to fit this rig. Additionally, the retargeting does not consider the wide shells on Cosmo. This is especially prevalent in the foot shells and results in the meshes clipping into each other.

B. Imitation Learning with AMP

We frame the locomotion problem as a Partially Observable Markov Decision Process (POMDP) where the agent must learn a policy $\pi(a|s)$ that maps observations s to actions a. The training objective maximizes the expected discounted return:

$$J(\pi) = \mathbb{E}_{\tau \sim \pi} \left[\sum_{t=0}^{T} \gamma^t r(s_t, a_t) \right] \tag{1}$$

Our approach utilizes Adversarial Motion Priors (AMP) [4] to learn locomotion from human motion capture clips. Figure 5 provides an overview of our complete training framework. AMP incorporates a discriminator network $D_\phi(s)$ that distinguishes motions from the reference dataset $\mathcal{M}$ from states generated by the policy π_θ, providing a learned reward signal that encourages lifelike motion.

The loss function for the discriminator is

$$\begin{aligned} \mathcal{L}_D(\phi) = &- \mathbb{E}_{s_t \sim \pi_\theta}[\log(1 - D_\phi(s_t))] \\ &- \mathbb{E}_{s_t^{ref} \sim \mathcal{M}}[\log(D_\phi(s_t^{ref}))]. \end{aligned} \tag{2}$$

The policy then incorporates the discriminator's output as a reward term $r_{AMP}(s_t) = \log D_\phi(s_t)$, encouraging the agent to generate motions that appear natural while simultaneously satisfying the task objectives and physical constraints. The discriminator is trained alongside the policy in an adversarial manner, continuously adapting to distinguish increasingly realistic policy behaviors. The policy network uses 3 layers (512, 256, 128 units), critic and discriminator use 2 layers (256, 128 units), all with ELU activations.

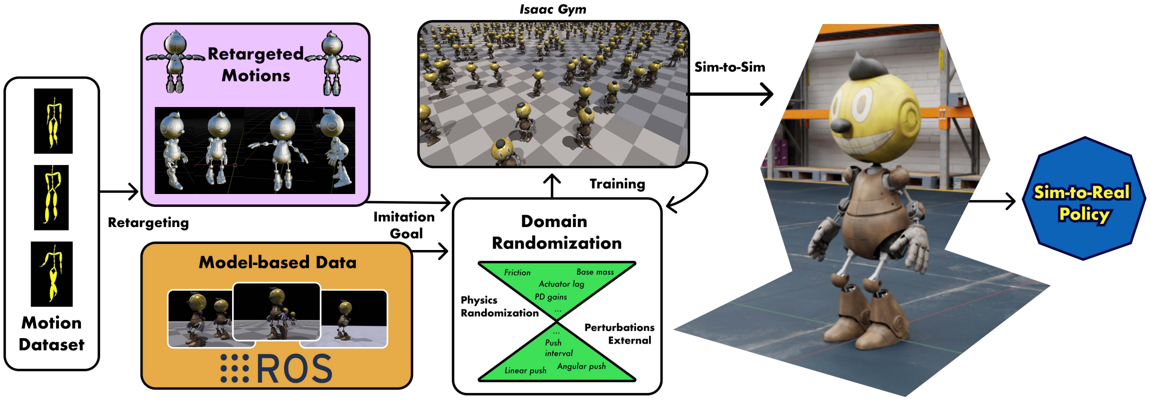


Fig. 3: Sim-to-Real pipeline: (a) Retargeting from diverse data sources, (b) Training, (c) Validation, (d) Deployment

Observation space Our observation is a state vector $s \in \mathbb{R}^{d_s}$ comprising proprioceptive information acquired from the motor chain and onboard state estimator:

$$s = [v_{\text{base}};\ \omega_{\text{base}};\ q - q_{\text{default}};\ \dot{q};\ g_{\text{proj}};\ a_{\text{prev}};\ h_{\text{base}};\ c_{\text{cmd}}].$$

This formulation captures the robot's complete dynamic state through base linear velocity ($v_{\text{base}} \in \mathbb{R}^3$), angular velocity ($\omega_{\text{base}} \in \mathbb{R}^3$), normalized joint positions ($q - q_{\text{default}} \in \mathbb{R}^{n_j}$), joint velocities ($\dot{q} \in \mathbb{R}^{n_j}$), projected gravity orientation ($g_{\text{proj}} \in \mathbb{R}^3$), previous actions ($a_{\text{prev}} \in \mathbb{R}^{n_j}$), base height ($h_{\text{base}} \in \mathbb{R}$), and command signals ($c_{\text{cmd}} \in \mathbb{R}^3$). Base kinematics are estimated via an invariant extended Kalman filter using proprioceptive data.

The command ranges in Table I reflect the typical operating range of Cosmo, with forward velocity having an asymmetric range to accommodate the natural bias toward forward locomotion.

TABLE I: Commanded velocity components in c_{cmd}

Description	Command Range	Units
Desired forward velocity	$[-0.3, 0.9]$	m/s
Desired lateral velocity	$[-0.3, 0.3]$	m/s
Desired yaw rate	$[-0.3, 0.3]$	rad/s

Action space The action space consists of target joint positions for all actuated degrees of freedom. These target positions are converted to torques through a PD controller running at a higher frequency than the policy.

Reward Scheme Our reward function combines task-oriented terms with the style reward from AMP. The task rewards are organized into distinct groups, as shown in Table II, with adjustable coefficients to enable curriculum learning.

We carefully designed reward groups to ensure motion quality and safety while accomplishing the locomotion task. The reward structure comprises three key functional areas and later will be used as metric to compare the performance of different experiments:

1) The *Motion Quality* rewards limit joint rate changes to ensure smooth rather than jerky movements. Joint target rates are computed as the L2 norm of consecutive joint target differences, scaled by the control frequency.

TABLE II: Reward Components for AMP Control

Components	Formula
Imitation	
AMP Reward	$-\log D_\phi(s)$
Motion Quality	
Joint rate	$\exp(-(\dot{q} - \dot{q}_{\text{target}})^2\sigma^2)$
Safety	
Foot stumble	$\exp(-(F_{\max} - F)^2/\sigma^2)$
Foot orientation	$r = \exp(-\lVert\vec{n}_{\text{feet}} - \vec{n}_{\text{ref}}\rVert^2/\sigma^2)$
Foot height	$r = \exp(-(h_{\text{feet}} - h_{\text{ref}})^2/\sigma^2)$
Task	
Linear velocity	$\exp(-(v_{\text{cmd}} - v_{\text{base}})^2/\sigma^2)$
Angular velocity	$\exp(-(\omega_{\text{cmd}} - \omega_{\text{base}})^2/\sigma^2)$

2) The *Safety* rewards protect the robot's aesthetic shells and optimize foot interactions from aggressive impacts. For example, our foot orientation reward specifically keeps feet flat to the plane, preventing the brown protective covers (Figure 4a) from contacting the floor at damaging angles.
3) The *Task Reward* components directly address the locomotion objectives.

Our collision modeling used simplified convex hulls to represent shells during training while accounting for motion constraints—particularly around the feet where clearance is minimal, and for internal motors that limit yaw rate due to proximity.

Ablation studies in Section V demonstrate each reward group's critical importance. This balanced reward approach, combined with our simulation-first methodology, minimized physical testing risks and identified potential issues before deployment.

C. Hardware Platform

Our custom humanoid platform, Cosmo, embodies a fictional character from a blockbuster movie while delivering robust locomotion capabilities. Cosmo's design presents unique control challenges with its disproportionately large head weighing 4 kg—16% of the 25 kg total mass—creating a top-heavy distribution rarely addressed in humanoid locomotion research (Figure 2).

The robot relies exclusively on proprioceptive feedback from its 28 degrees of freedom: 10 in the legs, 8 in the arms, 2 in the head, and 8 in the hands.

For locomotion control, Cosmo employs Westwood Robotics actuators—Panda Bear Plus models for high-torque hip pitch and knee joints, and Koala Bear Muscle Build variants for other primary joints—enabling precise torque control through internal sensing alone.

The design effectively balances aesthetic requirements with functional engineering within severely constrained packaging volumes. The character-defining large head incorporates carbon fiber reinforcement to minimize weight while maintaining visual presence. The feet include spring steel flexures, providing compliance to absorb impact from uneven terrain.

Strategic joint design optimizes mass distribution by positioning actuators close to the torso and minimizing limb inertia. These engineering solutions enable Cosmo to achieve stable locomotion despite its entertainment-first morphology with significantly elevated center of mass.

D. Simulation Suite

To systematically address the particular instability caused by Cosmo's head, we developed a comprehensive simulation environment in NVIDIA's Isaac Sim [18] before physical deployment. The physics engine enabled detailed analysis of both environmental and self-collisions, critical for evaluating stability with the disproportionately massive head. For stability assessment, we defined the center of mass, projected gravity vector, and kinematically feasible poses. We constructed a support polygon as the convex hull of foot-ground contact points, with stability determined by whether the gravity vector remained within this polygon during motion transitions. Our methodology involved systematically testing predefined poses by sweeping each joint through its range of motion while monitoring for collisions and stability violations.

We validated these poses on the physical platform to establish sim-to-real correspondence and determine the optimal stance for Cosmo. Comparative analysis against a biomechanically optimized robot ARTEMIS [19] quantitatively demonstrated Cosmo's inherent instability challenges, revealing center of mass deviations five times greater during equivalent motions.

After stability analysis, we conducted sim-to-sim transfer in Isaac Sim to validate policies safely before hardware deployment. This enabled joint gain tuning without endangering the physical robot and testing of controller robustness against external disturbances and varying surface friction.

For massive parallel training, we employ our pipeline in Isaac Gym [20], [21], NVIDIA's high-performance physics simulation platform.

Push-Recovery/ Balancing Policy Based on the pose stability analysis, we implemented an intermediate policy development stage to address two critical concerns: safeguarding the robot's delicate components and acquiring valuable practical insights for more complex motion control. The push-recovery policy served as an intuitive debugging case with well-defined expected behaviors.

For implementation, we utilized a concise motion clip featuring the statically stable pose identified in our stability analysis, maintained at a fixed position. We trained a push-recovery policy by applying random external forces, allowing the system to learn corrective actions.

This intermediate policy helped identify the most sensitive parameters of our platform—findings that were subsequently validated in our experimental results. Critically, this approach also facilitated efficient parameter tuning when transitioning between our two Cosmo platforms, as the push-recovery framework provided a controlled environment to calibrate platform-specific differences while minimizing risk to hardware.

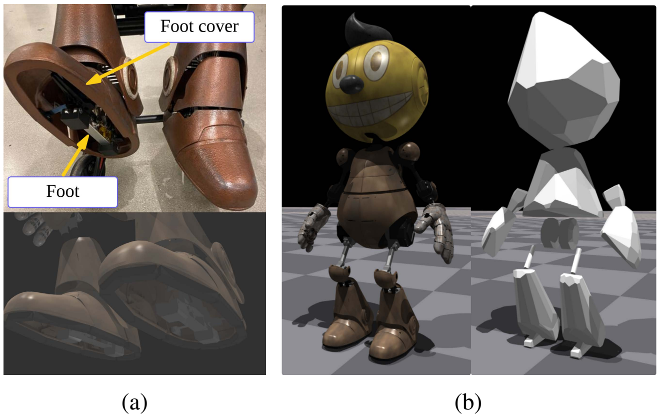


Fig. 4: (a) Real vs. simulated feet showing protective aesthetic covers. (b) Cosmo's collision meshes and foot model.

Walking Policy For our walking policy, we combined diverse motion references to achieve both expressiveness and stability. While we began with motion capture trajectories including ”swaggy” walking and standing poses, this data alone proved insufficient for safe deployment—lacking robot-specific dynamics information and offering only fixed velocities typical in mocap datasets. To address these limitations, we supplemented our references with experimental motion data from a hierarchical quadratic programming based inverse dynamics whole-body controller running on Cosmo (Figure 3), which provided the necessary dynamics information and a continuous spectrum of velocities as commanded. This hybrid approach created a natural curriculum, starting with controlled, lower-speed movements before progressing to more expressive gaits (typically 0.5-0.7 m/s), enabling safe hardware testing while preserving human-like expressiveness.

E. Sim-to-Real Transfer

Our hardware implementation employs low-level PD control for policy-to-actuation translation:

$$\tau = K_p(q_{\text{target}} - q) + K_d(\dot{q}_{\text{target}} - \dot{q})$$

We established a gain-tuning protocol focused on joint tracking, vibration management, and torque constraints, with

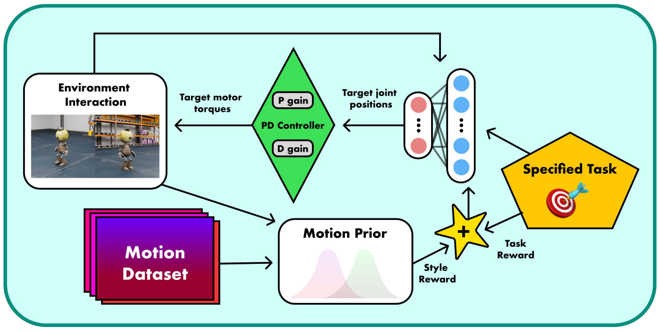


Fig. 5: Schematic overview of the training framework.

TABLE III: Domain Randomization Parameters

Parameter	**Description**	**Range**
Physics Randomization		
Friction	Contact friction coefficient	[0.2, 1.1]
Base mass	Robot base mass perturbation	±1.5 kg
PD gains	Joint PD controller gain multipliers	[0.75, 1.13]
Actuator lag	Control signal delay	4 timesteps
External Perturbations		
Push interval	Time between successive external forces	4.0 s
Linear push	Maximum linear push velocity	0.5 m/s
Angular push	Maximum angular push velocity	0.2 rad/s

more aggressive tuning for torso-related joints to manage the robot's disproportionate head mass.

To bridge the reality gap, we implemented domain randomization (Table III) through two key strategies: 1) dynamics randomization of physical parameters (mass, gains, friction, actuator response), and 2) calibrated sensor noise injection to develop policy robustness. We also incorporated periodic external perturbations with tuned magnitudes and intervals.

IV. Experimental Results

Despite the robot's challenging morphology and constrained range of motion, our trained policy demonstrates stable standing and walking on flat surfaces. In this section, we outline the primary challenges encountered during the training process, simulation, and hardware deployment, along with the strategies employed to address them. Our approach involved an extensive study of the platform using cutting-edge simulators followed by numerous hardware trials. A comprehensive justification of our design and methodological choices is provided in the ablation studies section, where we further analyze the specific difficulties encountered and the effectiveness of our solutions.

A. Pose Stability Analysis and range of motion test (Isaac Sim)

Static Analysis To identify stable configurations, we evaluated four poses derived from a standard human standing stance with arms at the sides. Using the projected gravity vector relative to the foot-defined stability region, we assessed pose viability during motion. In simulation, only 2 of 4 poses maintained stability throughout their joint range of motion, while the others resulted in falls when the gravity vector exited the support region. Additional results for tested poses are included in the paper's corresponding video.

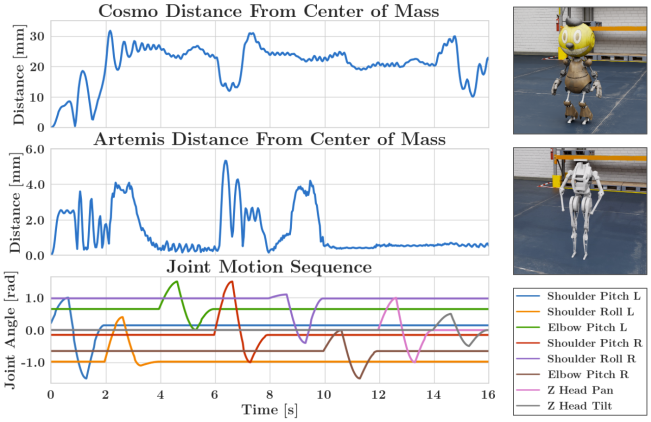


Fig. 6: Pose stability testing tracking the center mass disturbance from the initial center of mass position for stable poses. (**Top**) test results on Cosmo, (**Middle**) test results on a traditional humanoid ARTEMIS, (**Bottom**) the position of the all joints during test, causing center of mass disturbances. Both robots share identical arm configurations, so joint correspondences are equivalent.

Figure 6 illustrates this analysis and our selected initial pose. The top plot tracks the Euclidean distance of the center of mass from its initial position during joint motion in a stable pose. For comparison, the middle plot shows identical testing on ARTEMIS. The results are striking: Cosmo's center of mass deviated by 30 mm—five times more than ARTEMIS (6 mm)—quantitatively confirming our entertainment-focused design's inherent instability challenges. The bottom plot displays the joint motion sequence used in testing to cause the devitations. For example, the movement of the right shoulder pitch joint corresponds to a large disturbance in both models, which matches intuition that forward motion of the arm limb would cause disturbances to stability. This analysis directly informed our initial pose selection for control development and training, establishing a practical starting point for managing the robot's challenging mass distribution.f

B. Training results (Isaac Gym)

We identified optimal training parameters using Random State Initialization across different starting conditions. For performance evaluation, we used two key metrics: reward components (where values closer to 1.0 indicate better performance) and AMP discriminator loss (where values closer to 0 indicate better style matching with reference motions).

1) Balancing: Training proved highly sensitive to the robot's head mass distribution. We systematically evaluated balancing policy performance by varying head mass in the robot's URDF file while maintaining consistent inertial distribution and head randomization parameters, as shown in Table IV.

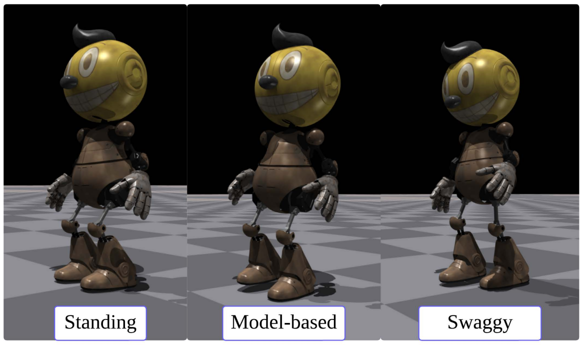


Fig. 7: AMP policies with styles for balancing, model-based walking and walking with swagger.

Our findings reveal a non-linear relationship between head mass and policy performance. The optimal head mass is 3.2 kg, achieving the best average performance (0.677), suggesting that moderate top-heaviness benefits the system by providing more pronounced proprioceptive signals while maintaining manageable inertial properties. Reducing mass to 2.2 kg presented challenges (0.660 average), as the controller struggled to develop a generalizable policy for lighter heads.

Increasing the head mass beyond 3.2 kg consistently degraded performance: 4.2 kg dropped to 0.658, 5.2 kg fell significantly to 0.613, and 6.2 kg recovered slightly to 0.653 but remained well below optimal. This demonstrates that excessive mass creates inertial challenges that overwhelm the policy's compensation capabilities.

2) Walking: As illustrated in Figure 7, we developed three distinct policy styles: a basic standing pose, a model-based walking gait, and a more dynamic "swaggy" style incorporating natural human-like motion. These variations demonstrate our AMP framework's flexibility in capturing diverse movement aesthetics while maintaining stability.

Our walking policy evaluation focused on two critical parameters: reward structure weights and AMP style coefficient. Each row in Table IV represents a separately trained policy under the specified configuration, not evaluation of a single policy under different conditions. For a fair comparison of styles, a single baseline discriminator is used to evaluate the style rewards across experiments. For reward structure, the balanced distribution [0.35, 0.35, 0.4] achieved the best average performance (0.721) with strong safety metrics (0.693). Alternative distributions like [0.5, 0.3, 0.2] improved individual task performance (0.757) but reduced safety (0.567).

For AMP style coefficient, a lower value (0.4) provided optimal overall performance (0.733) with excellent safety (0.720), while higher coefficients showed diminishing returns in safety: 0.7 achieved 0.729 average with 0.696 safety, and 0.9 reached 0.719 but degraded safety to 0.679. However, higher coefficients produced more human-like motion (discriminator loss improved from -0.420 to -0.372), highlighting the trade-off between robust performance and style fidelity for deployment.

TABLE IV: Performance metrics for balancing and walking tasks. Avg. represents the mean average of the previous 3 metrics. All results are averaged over 5 random seeds

Balancing - Added Mass					
kg	**Motion**	**Task**	**Safety**	**Avg.**	**AMP**
2.2	0.680±0.095	0.647±0.085	0.654±0.089	0.660±0.090	-
3.2	0.**693±0.091**	**0.663±0.085**	**0.675±0.090**	**0.677±0.089**	-
4.2	0.675±0.119	0.640±0.106	0.658±0.112	0.658±0.113	-
5.2	0.631±0.092	0.594±0.081	0.615±0.087	0.613±0.087	-
6.2	0.666±0.122	0.636±0.112	0.658±0.117	0.653±0.117	-
Walking - Reward Structure					
[.35, .35, .4]	0.723±0.107	0.743±0.118	0.693±0.109	**0.721±0.111**	-0.392±0.036
[.5, .3, .2]	**0.773±0.113**	**0.757±0.123**	0.567±0.091	0.699±0.110	**-0.366±0.039**
[.2, .6, .2]	0.710±0.108	0.758±0.118	0.566±0.090	0.678±0.106	-0.358±0.039
[.2, .4, .4]	0.678±0.102	0.739±0.117	**0.699±0.109**	0.703±0.110	-0.381±0.037
Walking - AMP Style Coefficient					
0.4	**0.726±0.109**	**0.754±0.119**	**0.720±0.111**	**0.733±0.113**	-0.420±0.029
0.5	0.722±0.109	0.749±0.118	0.715±0.112	0.720±0.113	-0.412±0.029
0.7	0.724±0.110	0.738±0.117	0.696±0.110	**0.729±0.112**	**-0.372±0.033**
0.9	0.735±0.114	0.742±0.118	0.679±0.109	0.719±0.114	-0.378±0.035

C. Sim-to-Real Transfer (Real Hardware)

Our AMP-trained policy achieves robust balancing capabilities as demonstrated in Figure 8. The upper body tracking data shows effective head stabilization despite its significant mass, with minimal deviations even during substantial perturbations up to 0.15 m/s.

Lower body tracking reveals sophisticated joint behaviors: Hip joints maintain stability while Ankle Pitch exhibits controlled oscillations during disturbance events (at 15s, 20s, and 30s), demonstrating that the policy has successfully learned human-like ankle-prioritized recovery strategies.

Body-local velocity data confirms strong disturbance rejection, with the system regaining stability within 2 seconds even after complex multi-directional perturbations. These results validate our domain randomization approach and reward structure.

The walking policy demonstrates successful transfer from simulation to hardware as shown in Figure 9. This natural walking implementation exhibits the dynamic, human-like characteristics that AMP is designed to produce. For instance, the policy reproduces key human movement patterns from the mocap data, including coordinated shoulder swing during stride and subtle head oscillations that create a more lifelike appearance. The data shows purposeful joint oscillations and pronounced velocity fluctuations in the x-direction, reaching peaks of 0.4 m/s commanded with distinctive patterns that authentically reproduce human walking gait cycles. This style prioritizes expressiveness, creating more engaging locomotion ideal for entertainment applications.

The policy successfully manages the robot's challenging mass distribution while maintaining joint torques within the safe operating range of ±20 Nm. Torque commands are clamped to actuator-specific torque limits, enforced in both simulation and hardware. The variable torque patterns during weight transfer phases reflect the sophisticated dynamics of this expressive gait.

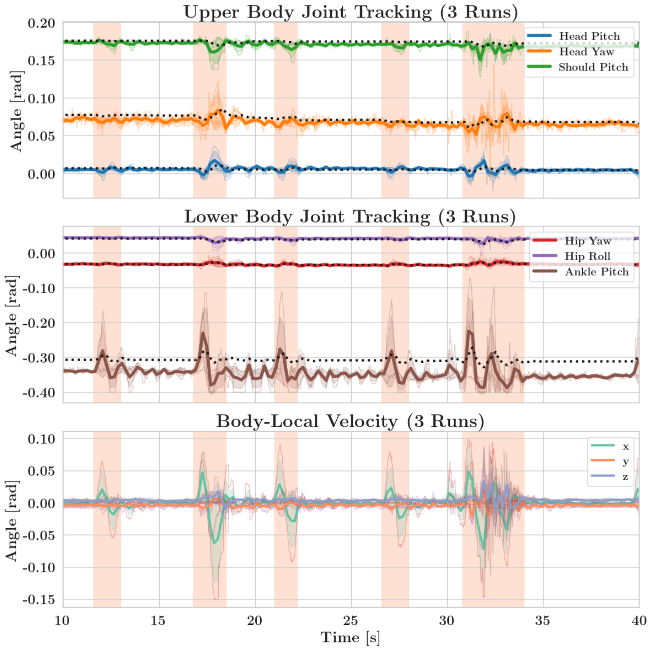


Fig. 8: Joint tracking and disturbance rejection. Shaded regions indicate disturbance periods.

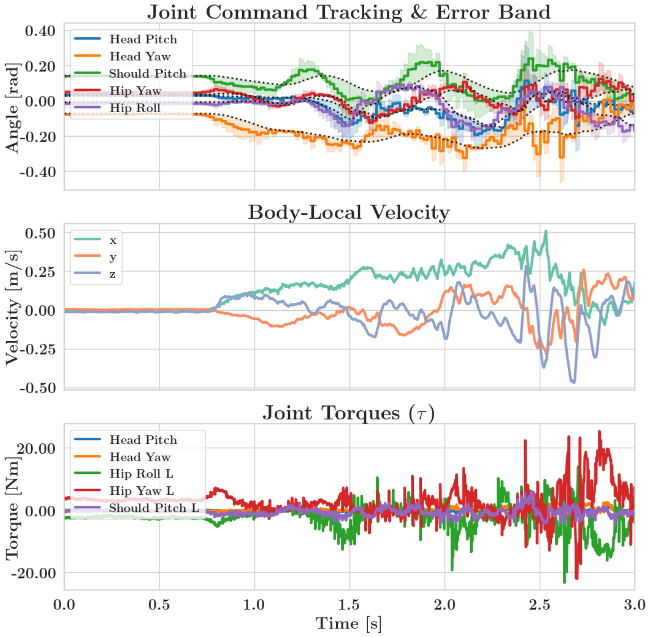


Fig. 9: Joint command tracking, body-local velocity and joint torque tracking for natural walking.

These results validate our sim-to-real approach and demonstrate AMP's effectiveness in adapting to entertainment-focused designs while preserving natural human-like motion. Notably, the policy successfully respects hardware limitations and shell integrity constraints despite Cosmo's challenging morphology. Our successful hardware implementation confirms that our comprehensive domain randomization strategy effectively bridges the reality gap, even for robots with such demanding aesthetic and structural constraints.

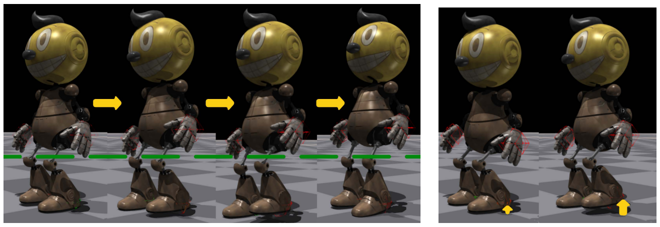
V. Ablation Studies

A. Motion Styles

Our systematic ablation studies conclusively demonstrate the necessity of combining multiple reference types for achieving stable, natural locomotion, as shown in Figure 10.

Without the standing pose reference, the robot exhibits dangerous vertical motion when transitioning to walking. Figure 10a reveals a nearly jumping behavior that significantly raises the center of mass above the reference line—dramatically increasing instability risk and energy consumption. This finding validates our methodical implementation approach from push recovery to a walking policy.

More significantly, Figure 10b demonstrates that removing the model-based reference produces ineffective high-frequency stepping with poor foot trajectory control. Without this reference data guiding low-velocity transitions, the policy fails to accelerate from standstill to human-like walking speeds smoothly. The resulting gait exhibits rapid, inefficient foot placement without proper leg elevation. Importantly, our reward-only approach proves insufficient—proper foot trajectories require explicit reference motions rather than reward-based compensation alone.



(a)

(b)

Fig. 10: Motion reference ablation: (a) Without standing reference; (b) Without model-based reference.

B. Reward Component Analysis

Our analysis establishes that specialized rewards beyond core imitation, motion quality, and task rewards are essential for managing foot trajectories and impact forces. Figure 11 compares three configurations: our baseline approach (all rewards), variants without stumble prevention, and without foot height rewards.

The baseline configuration achieves controlled foot trajectories with moderate contact forces distributed over longer periods. At timestep 45, a complete baseline step takes approximately 12 control ticks, enabling gentle foot placement and preventing dangerous air phases. This translates to a biologically plausible gait frequency of approximately 1.9 Hz (based on 26 ticks × 0.02s per tick), which aligns well with natural human walking patterns. In contrast, configurations missing either height or stumble rewards produce faster but hazardous motions with dangerous impact force

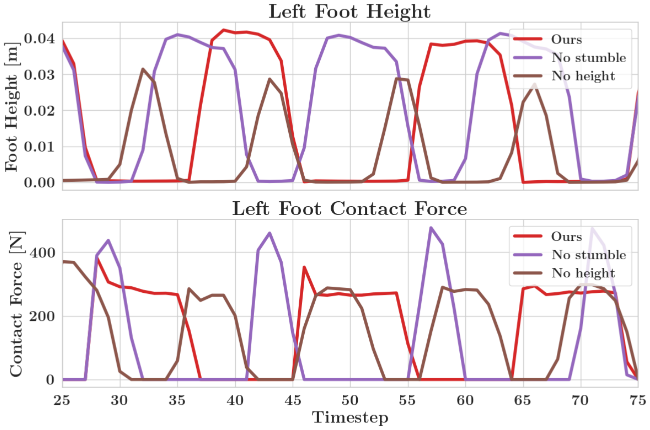


Fig. 11: Foot height and contact force comparison under different reward configurations, illustrating the necessity of reward engineering for safety precautions

spikes, creating potential risks for hardware damage during deployment.

These specialized rewards protect the robot's physical components from damaging torque spikes during operation, enabling safer walking without compromising mobility.

VI. Discussion and Conclusion

This work demonstrates the effectiveness of integrating physics-based simulation, adversarial motion priors, and domain randomization to enable locomotion in robots with challenging non-standard morphologies. Our successful implementation on Cosmo—with its disproportionately large head (16% of total weight), complex collision geometries, and restrictive foot shells—confirms that AMP-guided reinforcement learning can adapt to aesthetic design constraints that defeat traditional approaches. This work demonstrates the viability of the AMP framework with eccentric morphologies, demonstrating behavioral adaptations—conservative ankle strategies and increased stabilization—while maintaining human-like motion style.

Key findings include: (1) specialized rewards are essential for hardware protection; (2) multiple motion references provide expressiveness and stability; and (3) domain randomization bridges reality gaps for unusual morphologies. While our approach prioritizes stability, platform safety, and natural motion over speed (achieving 0.5-0.7 m/s gaits), this trade-off aligns well with entertainment applications where expressiveness and hardware preservation often matter more than agility.

Future work will compare imitation learning against traditional model-based controllers for entertainment robotics, where non-standard proportions and aesthetic constraints challenge conventional control approaches, and explore cross-embodiment transfer to validate broader applicability across diverse aesthetic morphologies.

References

[1] Z. Gu, J. Li, W. Shen, W. Yu, Z. Xie, S. McCrory, X. Cheng, A. Shamsah, R. Griffin, C. K. Liu, A. Kheddar, X. B. Peng, Y. Zhu, G. Shi, Q. Nguyen, G. Cheng, H. Gao, and Y. Zhao, "Humanoid locomotion and manipulation: Current progress and challenges in control, planning, and learning," 2025. [Online]. Available: https://arxiv.org/abs/2501.02116

[2] A. 360, "So lifelike! cutest toothless hiccup interaction ever at the new epic universe theme park," https://youtu.be/U6f3OrO-DlM, 2025, youTube video, published April, 2025.

[3] R. Grandia, E. Knoop, M. A. Hopkins, G. Wiedebach, J. Bishop, S. Pickles, D. Müller, and M. Bächer, "Design and control of a bipedal robotic character," *arXiv preprint arXiv:2501.05204*, 2025.

[4] X. B. Peng, Z. Ma, P. Abbeel, S. Levine, and A. Kanazawa, "Amp: adversarial motion priors for stylized physics-based character control," *ACM Transactions on Graphics*, vol. 40, no. 4, p. 1–20, Jul. 2021. [Online]. Available: http://dx.doi.org/10.1145/3450626.3459670

[5] K. Yamamoto, T. Kamioka, and T. Sugihara, "Survey on model-based biped motion control for humanoid robots," *Advanced Robotics*, vol. 34, no. 21-22, pp. 1353–1369, 2020.

[6] S. Kuindersma, R. Deits, M. Fallon, A. Valenzuela, H. Dai, F. Permenter, T. Koolen, P. Marion, and R. Tedrake, "Optimization-based locomotion planning, estimation, and control design for the atlas humanoid robot," *Autonomous Robots*, vol. 40, no. 3, pp. 429–455, 2016.

[7] 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. [Online]. Available: https://www.science.org/doi/abs/10.1126/scirobotics.adi9579

[8] Z. Li, X. B. Peng, P. Abbeel, S. Levine, G. Berseth, and K. Sreenath, "Reinforcement learning for versatile, dynamic, and robust bipedal locomotion control," *The International Journal of Robotics Research*, vol. 0, no. 0, 2024.

[9] X. Cheng, Y. Ji, J. Chen, R. Yang, G. Yang, and X. Wang, "Expressive whole-body control for humanoid robots," in *Robotics: Science and Systems*. Delft, Netherlands: RSS Foundation, 2024. [Online]. Available: https://www.roboticsproceedings.org/rss20/p107.html

[10] Z. Li, X. Cheng, X. B. Peng, P. Abbeel, S. Levine, G. Berseth, and K. Sreenath, "Reinforcement learning for robust parameterized locomotion control of bipedal robots," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 2811–2817.

[11] Z. Zhuang, S. Yao, and H. Zhao, "Humanoid parkour learning," *arXiv preprint arXiv:2406.10759*, 2024.

[12] C. Zhang, W. Xiao, T. He, and G. Shi, "Wococo: Learning whole-body humanoid control with sequential contacts," in *8th Annual Conference on Robot Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=Czs2xH9114

[13] 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)*. Kyoto, Japan: IEEE, 2022, pp. 1–8.

[14] Rokoko, "Rokoko studio live plugin for blender," https://github.com/Rokoko/rokoko-studio-live-blender, 2025, version 1.4.1, accessed April 28, 2025.

[15] Blender Online Community, "Blender – a 3d modelling and rendering package," Blender Foundation, Amsterdam, The Netherlands, 2025, version 4.4. [Online]. Available: https://www.blender.org

[16] C. M. University, "CMU Graphics Lab Motion Capture Database," http://mocap.cs.cmu.edu, 2003, accessed: 04/01/25.

[17] R. Milk, "Massive library of free 3d character animations," 2022. [Online]. Available: https://rancidmilk.itch.io/free-character-animations

[18] NVIDIA, "Isaac sim on omniverse," 2023, https://developer.nvidia.com/isaac-sim.

[19] T. Zhu, "Design of a highly dynamic humanoid robot." [Online]. Available: https://escholarship.org/uc/item/0qz3p57g

[20] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, and G. State, "Isaac gym: High performance gpu-based physics simulation for robot learning," 2021. [Online]. Available: https://arxiv.org/abs/2108.10470

[21] N. Rudin, D. Hoeller, P. Reist, and M. Hutter, "Learning to walk in minutes using massively parallel deep reinforcement learning," *CoRR*, vol. abs/2109.11978, 2021. [Online]. Available: https://arxiv.org/abs/2109.11978