

PolySim: Bridging the Sim-to-Real Gap for Humanoid Control via Multi-Simulator Dynamics Randomization

Zixing Lei^{1,2,*}, Zibo Zhou^{1,*}, Sheng Yin^{1,*}, Yueru Chen^{1,*}, Qingyao Xu¹, Weixin Li², Yunhong Wang², Bowei Tang¹, Wei Jing³, and Siheng Chen^{1,3}

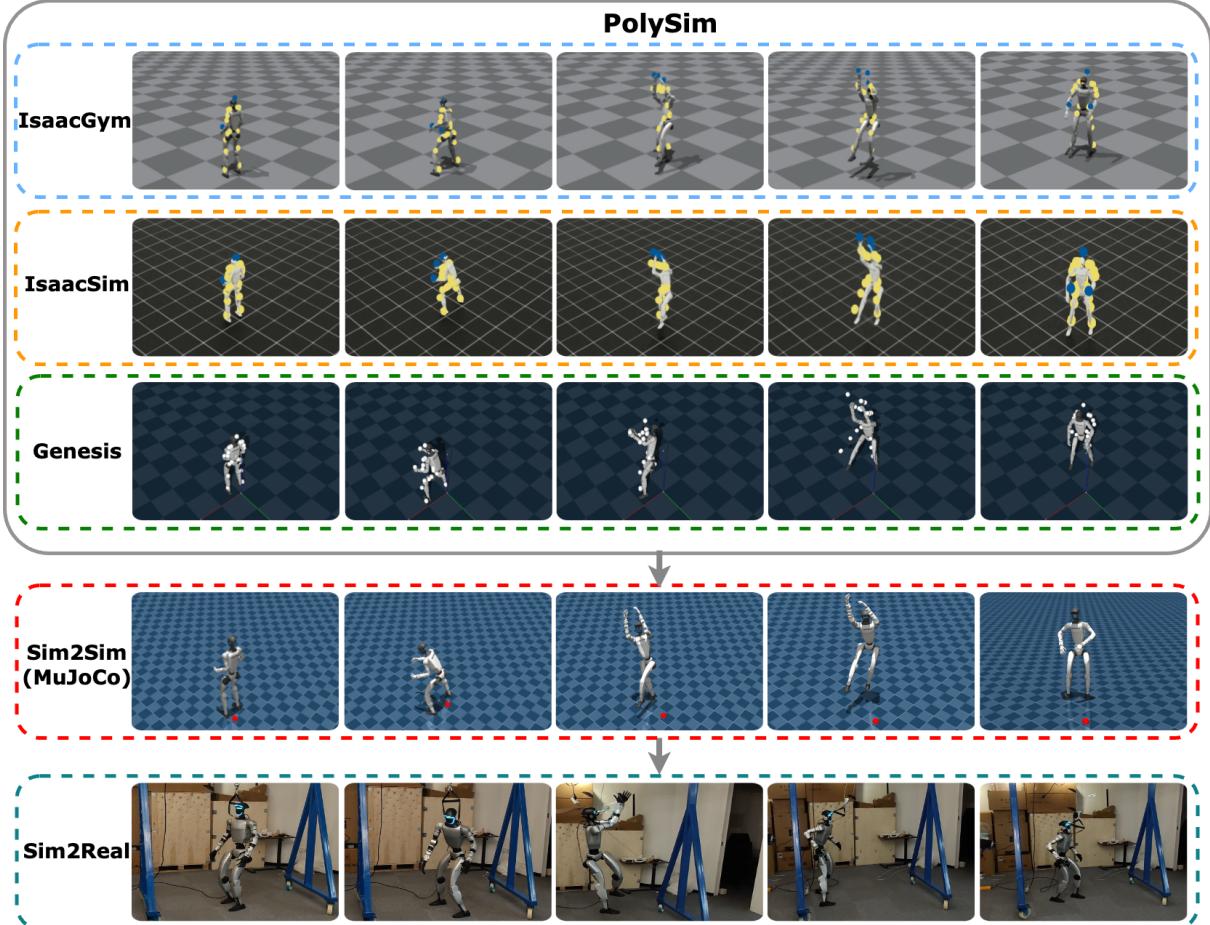


Fig. 1: **PolySim**, a parallel training framework, achieves whole-body agility on the Unitree G1 humanoid. Training across diverse simulators reduces motion-tracking error in sim-to-sim transfer and enables zero-shot deployment on the real world.

Abstract—Humanoid whole-body control (WBC) policies trained in simulation often suffer from the sim-to-real gap, which fundamentally arises from simulator inductive bias—the inherent assumptions and limitations of any single simulator. These biases lead to nontrivial discrepancies both across simulators and between simulation and the real world. To mitigate the effect of simulator inductive bias, the key idea is to train policies jointly across multiple simulators, encouraging the learned controller to capture dynamics that generalize beyond any single simulator’s assumptions. We thus introduce PolySim, a WBC training platform that integrates multiple heterogeneous simulators. PolySim can launch parallel environments from different engines simultaneously within a single training run,

thereby realizing *dynamics-level* domain randomization. Theoretically, we show that PolySim yields a tighter upper bound on simulator inductive bias than single-simulator training. In the experiments, PolySim substantially reduces motion-tracking error in sim-to-sim evaluations; for example, on MuJoCo, it improves execution success by 52.8% over an IsaacSim baseline. PolySim further enables *zero-shot* deployment on a real Unitree G1 without additional fine-tuning, showing effective transfer from simulation to the real world. The code is available at <https://github.com/EmboMaster/PolySim>

I. INTRODUCTION

Thanks to rapid progress in reinforcement learning (RL) and high-fidelity simulation, humanoid whole-body control (WBC) [1] has advanced markedly in recent years. Humanoid control is inherently high-dimensional and contact-rich, making the collection of sufficiently diverse real-world

*Equally contributed.

¹The authors from Shanghai Jiao Tong University.

²The authors from Zhongguancun Academy.

³The authors from Nerv.ai.

interaction data expensive and challenging. Consequently, simulator-based training provides a practical and scalable alternative. Simulation-based trajectory sampling circumvents the safety and efficiency limitations of real-world exploration, enabling the generation of massive rollouts.

Despite their utility, simulators provide only approximate models of real-world physics, each encoding its own inherent assumptions. Here we define simulator inductive bias as the inherent modeling assumptions of any simulator that determine its transition dynamics. As shown in Fig. 2, every simulator occupies a distinct position in the space of dynamics, while the real world lies elsewhere, with its own intrinsic bias. Policies trained in a particular simulator inevitably inherit its bias, hindering generalization across simulators and to the real world, and causing a significant sim-to-real gap.

To mitigate such a gap, the most common practice is domain randomization [2], [3], which perturbs observations, actions, latent states, or a limited set of physical parameters around a single simulator. This can partially improve generalization by exposing the policy to trajectories slightly beyond the nominal simulator dynamics. Nevertheless, even extensive randomization is structurally limited: the transition model remains that of the chosen simulator, determined by its modeling choices, contact solver, time integrator, and actuator models. Consequently, the randomized rollouts still occupy a narrow neighborhood of that simulator’s dynamics as shown in Fig. 2, causing the the sim-to-real gap still substantial.

To address this limitation, our key idea is to train policies jointly across multiple simulators, encouraging the learned controller to capture dynamics that generalize beyond any single simulator’s assumptions. Therefore, we propose **PolySim**, a WBC training platform that integrates multiple heterogeneous embodied simulators. **PolySim** enables domain randomization at the level of simulator dynamics, substantially mitigating the effect of any single simulator’s inductive bias. Compared with prior work [4], [5], **PolySim** provides three novel features: (i) training–simulation isolation via a client–server architecture that decouples RL training from simulator runtimes and supports flexible distributed execution; (ii) a unified simulator router that performs API translation and resource scheduling, enabling a single training loop to launch parallel environments backed by different simulators; and (iii) GPU pass-through communication that exchanges trajectories and diagnostics directly between training and simulation processes with negligible additional latency. Taken together, these features enable parallel, efficient sampling of diverse trajectories across heterogeneous simulator dynamics during training as shown in Fig 1, thereby improving generalization performance.

To validate the effectiveness of **PolySim**, we analyze its performance both theoretically and empirically. Theoretically, our analysis shows that the proposed **PolySim**, which performs dynamics-level randomization, admits a tighter upper bound on simulator inductive bias than approaches that only randomize at the parameter level. Empirically, our

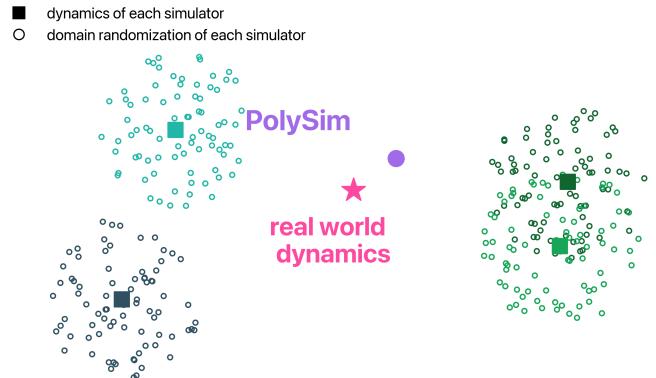


Fig. 2: Visual illustration of **PolySim**. The pink star denotes real-world dynamics. Filled squares indicate the nominal transition dynamics of each simulator, representing its inherent inductive bias as an approximation of real-world dynamics. Hollow circles depict domain-randomized variants that perturb parameters but remain centered around their respective simulators. Training against mixtures of simulators (**PolySim**, purple dot) combines multiple approximations of real-world dynamics, allowing the resulting policy to lie closer to the true dynamics than any single simulator or its parameter-level domain randomization, thereby reducing the notorious sim-to-real gap.

extensive experiments show that i) the proposed **PolySim** significantly reduces motion-tracking error compared to single-simulator baselines in sim-to-sim evaluations. For example, integrating IsaacSim, IsaacGym, and Genesis increases the motion-tracking execution success rate by **52.8%** over an IsaacSim-only baseline in MuJoCo; and ii) **PolySim** enables *zero-shot* deployment on the real Unitree G1 without additional fine-tuning, reflecting its effectiveness in both simulation and real-world deployment.

II. RELATED WORK

A. Learning-Based Methods for Humanoid Control

Learning-based whole-body control (WBC) for humanoids has rapidly advanced with deep RL [6] trained in modern physics simulators [7]–[10]. Recent systems demonstrate broad skill coverage: locomotion [11]–[13] on uneven or cluttered terrains [14], [15], as well as highly dynamic behaviors including stand-up [16], jumping [17], parkour [18], [19], and dancing [20]–[22]. In parallel, physics-based motion imitation continues to push agility and expressivity using human demonstrations and priors [23]–[25]. Despite these gains, most controllers are trained in a single simulator.

B. Sim-to-Real Methods for Robot RL

Bridging the reality gap typically follows two lines. The first is domain randomization (DR) within one simulator [3], [26]–[28], perturbing rendering and physics parameters to improve robustness; this helps but usually preserves the engine’s underlying transition model, leaving model-class mismatch unaddressed. The second is system identification (SysID) [29]–[31], which tunes simulators toward hardware using real data. Offline SysID [26], [32]–[36] calibrates

dynamics before policy learning, while online/adaptive variants [37]–[43] estimate latent properties or adapt policies at deployment. These strategies improve transfer yet often require substantial robot data, can be device-specific, and still inherit the bias of a single engine’s physics. We instead introduce a training-time alternative: parallel exposure to multiple simulators so that the policy optimizes against a mixture of transition models, thereby reducing simulator-induced bias while limiting real-data dependence.

C. Cross-Simulator Frameworks

HumanoidVerse modularizes simulators, tasks, and algorithms to ease switching among engines such as Isaac-Gym/IsaacSim/Genesis for sim2sim and sim2real studies [5]. *RoboVerse* introduces MetaSim, a universal interface that unifies heterogeneous engines and ships with a dataset and standardized benchmarks for IL/RL [4]. Both emphasize cross-simulator interoperability, but neither realizes *parallel* cross-environment RL that optimizes a single policy *concurrently* across simulators within one loop.

III. METHOD

To mitigate simulator-induced inductive bias, we operationalize transition-level domain randomization by optimizing a single policy against a mixture of heterogeneous dynamics within each training iteration. **PolySim** realizes this setting by orchestrating parallel rollouts across multiple simulators while exposing a simulator-agnostic interface to the learner (Fig. 3). Concretely, **PolySim** comprises three system designs: (i) **training–simulation isolation**, which decouples the RL learner from per-engine workers to enable distributed execution and failure isolation; (ii) a **simulator router** that harmonizes physics, translates APIs, and normalizes numerics so that all backends present identical observation–action–reward semantics; and (iii) **GPU pass-through** communication (RPC over NCCL) that keeps data on device for high-throughput parallel training.

A. Training–Simulation Isolation

Most existing WBC training stacks adopt a monolithic, centralized design: a single process both instantiates environments and updates the policy. While convenient for development, this conflates simulator and learner runtimes, causing (i) mutual dependency lock in simulation libraries and drivers constrain the training stack and (ii) hard resource contention simulation stepping and backpropagation compete for the same GPU budget.

We therefore execute training and simulation as *isolated processes*: a *TrainClient* and a set of *SimServers*. Each *SimServer* advances its local physics and returns typed payloads (o, r, d, info); the *TrainClient* aggregates trajectories and updates the policy. This separation (1) enables elastic, highly distributed execution across multiple GPUs, (2) preserves toolchain independence so the learner can adopt newer, faster software stacks without being constrained by simulator dependencies, and (3) provides strong fault isolation—failures or stragglers in one engine neither propagate

nor block others—allowing heterogeneous simulators to start and progress concurrently without resource interference.

B. Simulator Router — Unified Interface Virtualization

Heterogeneous simulators differ in scene parameterization, API surfaces, and numerical conventions, which prevents a single learner from receiving consistent inputs. Our simulator router virtualizes multiple simulators as one vectorized environment with invariant semantics, synchronized clocks, and device-resident tensors. As shown in Fig. 3, simulator router establishes a unified standard along three axes: (i) physics harmonization at initialization to construct maximally aligned scenes; (ii) API translation that exposes a common interface and delivers identically shaped observations and reward signals to the *TrainClient*; and (iii) numerical normalization to ensure that neural-network actions are correctly interpreted by each engine. With these alignments, SR abstracts away backend idiosyncrasies and enables stable, high-throughput parallel rollouts from heterogeneous simulators within a single training loop.

Physics Harmonization. Simulator router takes a unified physical specification and instantiates per-simulator scenes that are as consistent as possible. It maps friction and contact settings, actuator models and controller gains, gravity and integration step, rigid-body properties, and robot descriptions to the corresponding options of each simulator. Residual mismatches that arise from different contact solvers or integrators are recorded as fixed engine meta-parameters to provide controlled dynamics variation rather than uncontrolled drift.

API Translation. Simulator router exposes one environment interface to the learner by translating heterogeneous simulator APIs into a common observation–action–reward contract. It aligns state layouts, reference frames, and materializes only the physical quantities required by the policy and the reward. The result is identically shaped tensors with consistent meaning, independent of the underlying engine.

Numerical Normalization. Simulator router reconciles engine-specific numerical conventions so that neural-network outputs are applied correctly across backends. It performs unit conversion at the interface boundary such as standardizing angular quantities to radians, applies range mapping from a normalized action space such as rescaling $[-1, 1]$ actions to per-engine joint limits or position-target ranges, and enforces clipping consistent with actuator and controller limits to guarantee safety and correctness.

As a result, from the RL framework’s perspective all simulators deliver the same service: a unified interface with identical tensor shapes, semantics, and timing, while simulator router handles backend-specific details and routing.

C. GPU-Direct Communication Pipeline

To sustain high-throughput parallel rollouts across heterogeneous engines, **PolySim** employs a RPC pipeline over NCCL. Observations, actions, rewards, and masks are kept as device tensors end to end; cross-process exchanges use NVLink or high-speed PCIe without host staging, which removes CPU round trips and redundant serialization, and

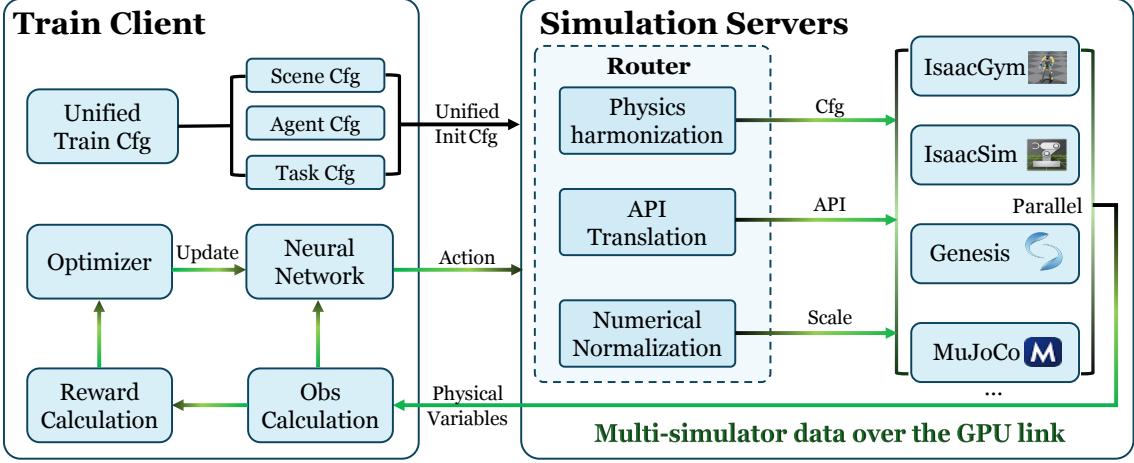


Fig. 3: System overview of the proposed parallel multi-simulator RL framework (Mode III). **Left (Training Framework):** a simulator-agnostic RL loop where a unified training configuration (scene/agent/task) drives observation and reward computation; the policy network produces actions and is updated by the optimizer. **Right (Simulation):** heterogeneous engines (IsaacGym/IsaacSim/Genesis/MuJoCo) are virtualized behind a *Simulator Router* that performs *physics harmonization*, *API translation*, and *numerical normalization*. The router maps the unified initialization config to engine-specific settings, dispatches actions, and returns physical variables for observation/reward calculation. Green paths indicate GPU-direct links (PyTorch RPC/NCCL over NVLink/PCIe), enabling concurrent rollouts across devices and stable, high-throughput training.

preserves bandwidth at RL rollout scales. This capability fully leverages CUDA RPC's ability to transmit tensors directly from local GPU memory to remote GPU memory and on NCCL's peer-to-peer transports.

IV. THEORETICAL SUPERIORITY OF POLYSIM

Setup and notation. Let the real environment be $M^* = (\mathcal{S}_0, \mathcal{A}, T_0, R_0, \gamma)$ and each simulator $M_i = (\mathcal{S}_i, \mathcal{A}, T_i, R_i, \gamma)$ be a projection via $f_i : \mathcal{S}_0 \rightarrow \mathcal{S}_i$, where \mathcal{S}_0 is the real state space, \mathcal{A} the common action space, $T_0(\cdot | s, a)$ the transition kernel on \mathcal{S}_0 , and R_0 a bounded reward; for simulator i , \mathcal{S}_i is its state space with transition kernel T_i and reward R_i . Given a measurable section $g_i : \mathcal{S}_i \rightarrow \mathcal{S}_0$ satisfying $f_i \circ g_i = \text{id}_{\mathcal{S}_i}$, define the lifted kernel

$$\tilde{T}_i(\cdot | s, a) := (g_i)_\# T_i(\cdot | f_i(s), a),$$

so that all kernels act on the common space \mathcal{S}_0 . For any policy π , let $\eta_{T,R}(\pi) := \mathbb{E}[\sum_{t \geq 0} \gamma^t R(s_t, a_t)]$ denote the discounted return under (T, R) . On (\mathcal{S}_0, d) , let W_1 denote the *1-Wasserstein distance induced by d* . Assume that for any π , the value function $V^{\pi,0}$ is L_V -Lipschitz on (\mathcal{S}_0, d) .

Definition 1 (Sim-to-real gap of a simulator): Let P_i be a simulator with \tilde{T}_i the corresponding lifted kernel. Then, the sim-to-real gap of P_i under a policy class Π is,

$$G_{S2R}(P_i; \Pi) \triangleq \sup_{\pi \in \Pi} |\eta_{\tilde{T}_i}(\pi) - \eta_{T_0}(\pi)|.$$

Lemma 1 (Upper-bound of the sim-to-real gap): The sim-to-real gap under a policy class Π is upper-bounded as

$$G_{S2R}(P_i; \Pi) \leq \frac{\gamma}{1-\gamma} L_V \Delta_i,$$

where $\gamma \in (0, 1)$ is the discount factor, L_V is the Lipschitz constant of $V^{\pi,0}$ and $\Delta_i := \sup_{s,a} W_1(T_0(\cdot|s, a), \tilde{T}_i(\cdot|s, a))$

Definition 2 (Sim-to-real gap of PolySim): Let Q be **PolySim**, a mixture of N simulators, whose transition kernel is

$$T_Q(\cdot | s, a) = \sum_{i=1}^N w_i \tilde{T}_i(\cdot | s, a) \in \text{conv}(\mathcal{K}),$$

where $w \in \Delta^{N-1}$ be the weight vector (the probability simplex) and $\mathcal{K} = \{\tilde{T}_1, \dots, \tilde{T}_N\}$ be the lifted simulator transition kernels on \mathcal{S}^0 . Then, the sim-to-real gap of Q under a policy class Π is

$$G_{S2R}(Q; \Pi) \triangleq \sup_{\pi \in \Pi} |\eta_{T_0, R_0}(\pi) - \eta_{\tilde{T}_Q, R_0}(\pi)|,$$

where $\eta_{T,R}(\pi)$ denotes the discounted return under (T, R) .

Theorem 1 (Superiority of the PolySim S2R Gap): **PolySim** achieves a lower sim-to-real gap than any single simulator. Given the lifted simulator kernels $\mathcal{K} = \{P_i\}_{i=1}^N$ on \mathcal{S}^0 (where $P_i := \tilde{T}_i$) and a convex mixture For any policy class Π , we have

$$G_{S2R}(Q; \Pi) \leq \min_i G_{S2R}(P_i; \Pi).$$

Theorem 1 shows that PolySim is better than any single simulator in terms of the sim-to-real gap. The intuition is that mixtures enlarge the feasible set from isolated simulators to their convex hull: when the real dynamics T_0 are strictly closer (in W_1) to $\text{conv}(\mathcal{K})$ than to any vertex P_i , hence the sim-to-real gap shrinks.

Proof: For any policy π and any kernel Q on \mathcal{S}_0 , writing the Bellman equations under (T_0, R_0) and (Q, R_0) , subtracting, taking $\|\cdot\|_\infty$, and using the L_V -Lipschitzness of $V^{\pi,0}$ together with the Kantorovich–Rubinstein dual for W_1 yields

$$\|V^{\pi,*} - V^{\pi,Q}\|_\infty \leq C \cdot \sup_{s,a} W_1(T_0(\cdot|s, a), Q(\cdot|s, a)).$$

Evaluating returns under any start-state distribution and taking absolute values gives the general value-gap bound

$$|\eta_{T_0, R_0}(\pi) - \eta_{Q, R_0}(\pi)| \leq C \cdot \sup_{s,a} W_1(T_0(\cdot|s, a), Q(\cdot|s, a)). \quad (1)$$

Taking $\sup_{\pi \in \Pi}$ on both sides, we obtain

$$G_{S2R}(Q; \Pi) \leq C \cdot \sup_{s,a} W_1(T_0(\cdot|s, a), Q(\cdot|s, a)). \quad (2)$$

Instantiate (2) with $Q = \tilde{T}_i$ and denote $\Delta_i := \sup_{s,a} W_1(T_0(\cdot|s, a), \tilde{T}_i(\cdot|s, a))$ to get

$$G_{S2R}(i; \Pi) \leq C \Delta_i,$$

which is exactly the claimed single-simulator upper bound. Let $\varepsilon_{\text{hull}} := \inf_{Q \in \text{conv}(\mathcal{K})} \sup_{s,a} W_1(T_0(\cdot|s, a), Q(\cdot|s, a))$ and take $Q \in \arg \min_{Q \in \text{conv}(\mathcal{K})} \sup_{s,a} W_1(T_0, Q)$. Applying (2) to Q gives $G_{S2R}(Q; \Pi) \leq C \varepsilon_{\text{hull}}$, while applying it to each \tilde{T}_i gives $G_{S2R}(i; \Pi) \leq C \Delta_i$. Under the stated assumption $\varepsilon_{\text{hull}} < \min_i \Delta_i$, there exists $\delta > 0$ s.t. $C \varepsilon_{\text{hull}} \leq C(\min_i \Delta_i - \delta)$. By the definition of the supremum and the tightness of the KR-dual bound (standard in communicating MDPs with stationary randomized policies), there exists a sequence $\{\pi_k\} \subset \Pi$ with $G_{S2R}(i^*; \Pi) \geq C \Delta_{i^*} - o_k(1)$ for some $i^* \in \arg \min_i \Delta_i$. Hence, for all sufficiently large k ,

$$\begin{aligned} G_{S2R}(Q; \Pi) &< C \Delta_{i^*} - o_k(1) \leq \\ \min_i G_{S2R}(i; \Pi) &\leq G_{S2R}(i^*; \Pi), \end{aligned}$$

which proves Lemma 1 and Theorem 1. The intuition is geometric: the true world dynamics (T_0) and each simulator kernel (\tilde{T}_i) are points in a metric space. PolySim’s mixed kernel (Q) can occupy any point within their convex hull, which we assume is strictly closer to T_0 than any single vertex. The key steps are: (1) Bounding the sim-to-real gap by the 1-Wasserstein distance, which connects the gap to this geometric distance. (2) Applying our assumption that the mixture’s minimal distance to reality ($\varepsilon_{\text{hull}}$) is strictly smaller than the best single simulator’s distance ($\min_i \Delta_i$), which proves PolySim achieves a tighter theoretical gap.

V. EXPERIMENT

In this section, we present experiments to evaluate the robustness of PolySim. Our experiments aim to answer the follow three key questions:

- Q1: Can PolySim outperform the policy trained on a single simulator to compensate for the dynamics mismatch?
- Q2: Can PolySim outperform policies trained sequentially across simulators?
- Q3: Does PolySim work for sim-to-real transfer?

A. Experiments Setup and Metrics

1) *Simulation Experiments*: We train policies on 14 motions of varying difficulty (easy, medium, hard) selected from the ASAP dataset [44], using their settings to train for 10k, 15k, and 20k iterations, respectively. The training is performed in parallel across a diverse set of simulators, including IsaacGym, IsaacSim, and Genesis. To purely evaluate the generalization benefit of PolySim, these policies

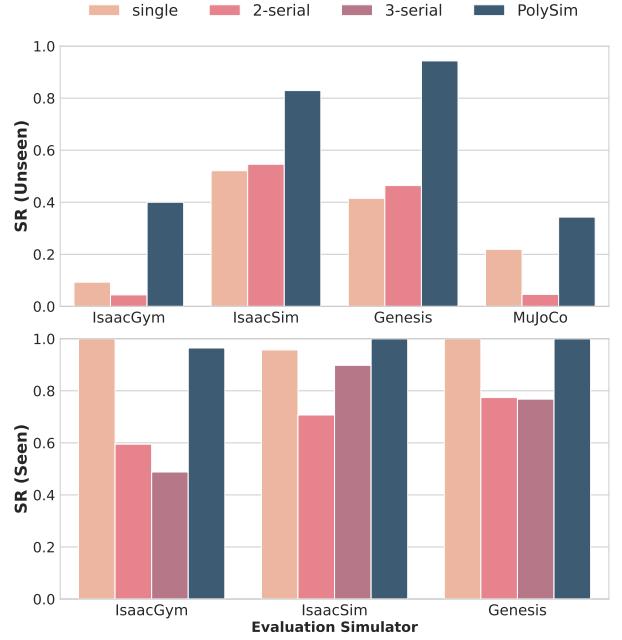


Fig. 4: Success rate on seen and unseen simulators under different settings. ‘single’ indicates training on a single simulator; ‘n-serial’ indicates sequential training on n simulators.

are then tested via zero-shot transfer to an entirely unseen simulator (MuJoCo) without applying any parameter-level domain randomization.

2) *Real-World Experiments*: For real-world evaluation, we deploy the trained policies on a Unitree G1 humanoid robot and execute motion sequences with substantial sim-to-real discrepancies to evaluate PolySim’s effectiveness in bridging the gap between simulation and physical systems. Specifically, we selected some challenging motions, which span a wide range of motor skills and showcase PolySim’s capability for agile whole-body motion control in real-world scenarios.

3) *Metrics*: To evaluate motion imitation and transfer performance, we adopt the following metrics: Success Rate: An imitation is deemed unsuccessful if at any point, the mean body position error exceeds 0.5 m. $E_{g-mpjpe}$ (mm): Global body position tracking error. E_{mpjpe} (mm): Root-relative mean per-joint position error. E_{acc} (mm/frame²): Acceleration error. E_{vel} (mm/frame): Root velocity error. All metrics are computed against the reference motion sequences, and mean values are reported across all sequences.

B. Baselines

We compare PolySim against the following baselines:

- **Single-Domain Training**: Policies are trained independently in each simulator, i.e., IsaacGym, IsaacSim, and Genesis, without heterogeneous domain parallelization.
- **Sequential Multi-Simulator Training**: Policies are trained on multiple simulators sequentially, where two or three simulators are selected in each experiment and executed in different orders for the same number of iterations. Each subsequent simulator inherits the policy and curriculum learning parameters from the previous stage.

TABLE I: Closed-loop motion imitation across test environments. Cells with backgrounds indicate *unseen* settings, where the test simulator was not included in training simulators. **Bold black** denotes the best outcomes under unseen settings.

Training Simulators	IsaacGym					IsaacSim				
	Succ ↑	$E_{g-mpjpe} \downarrow$	$E_{mpjpe} \downarrow$	$E_{acc} \downarrow$	$E_{vel} \downarrow$	Succ ↑	$E_{g-mpjpe} \downarrow$	$E_{mpjpe} \downarrow$	$E_{acc} \downarrow$	$E_{vel} \downarrow$
Genesis	0.043	201.570	97.421	4.492	13.745	0.721	150.560	80.380	6.112	9.018
IsaacGym	1.000	140.186	49.988	4.535	7.212	0.321	236.086	97.352	5.156	12.381
IsaacSim	0.143	189.434	83.321	4.466	13.208	0.957	121.407	44.237	3.549	6.196
IsaacSim+Genesis	0.400	159.611	74.142	4.400	10.608	1.000	102.065	49.506	3.376	6.324
IsaacGym+Genesis	1.000	108.971	51.898	3.624	6.380	0.829	181.457	83.016	5.939	9.988
IsaacSim+IsaacGym	1.000	113.443	47.038	3.773	6.505	1.000	114.100	45.428	3.548	6.185
IsaacSim+IsaacGym+Genesis	0.929	116.141	54.730	3.794	7.049	1.000	107.949	51.291	3.536	6.459

Training Simulators	Genesis					MuJoCo				
	Succ ↑	$E_{g-mpjpe} \downarrow$	$E_{mpjpe} \downarrow$	$E_{acc} \downarrow$	$E_{vel} \downarrow$	Succ ↑	$E_{g-mpjpe} \downarrow$	$E_{mpjpe} \downarrow$	$E_{acc} \downarrow$	$E_{vel} \downarrow$
Genesis	1.000	90.545	61.549	2.974	7.495	0.121	175.671	87.237	3.888	11.088
IsaacGym	0.229	105.378	72.596	4.661	10.136	0.500	214.159	71.286	6.191	11.910
IsaacSim	0.600	98.852	57.019	3.731	8.348	0.036	174.838	80.567	3.361	10.911
IsaacSim+Genesis	1.000	97.953	62.186	3.882	8.033	0.100	174.846	78.886	3.650	11.094
IsaacGym+Genesis	1.000	96.060	60.939	2.990	6.617	0.400	169.467	74.854	4.922	10.214
IsaacSim+IsaacGym	0.943	87.784	55.801	4.275	9.826	0.429	174.978	71.628	4.690	10.864
IsaacSim+IsaacGym+Genesis	1.000	97.728	64.879	3.876	8.437	0.564	151.924	65.594	4.308	9.329

C. Main Result

To address **Q1**, we conduct parallel training with diverse combinations of simulators. We then compare the resulting policies against those trained on a single simulator. Table I reports the performance of **PolySim** across different test environments. The results highlight several key findings:

1) *Superior generalization to unseen domains*: The most significant advantage of the PolySim strategy lies in its powerful zero-shot generalization capability, the ability to perform tasks in environments not encountered during training. MuJoCo, as an entirely unseen test domain, serves as the ultimate benchmark for evaluating a policy.

In the unseen MuJoCo domain, the Sim+Gym+Genesis configuration, leveraging the most diverse dynamics, achieved the best overall performance. It demonstrated superior robustness—with success rates increasing by 44.3%, 6.4%, and 52.8% over single-simulator baselines—and superior precision, with the lowest errors across key metrics ($E_{g-mpjpe}$, E_{mpjpe} , E_{vel}). This performance not only surpasses naive single-simulator training but also more advanced parameter-level DR (Table III). This confirms our central claim: true generalization requires overcoming a single engine’s inductive bias, a feat superficial randomization alone cannot achieve.

2) *Beyond additive effects: amplified gains from multi-simulator parallel training*: The improvement **PolySim** is not a simple additive effect of single-simulator models, but rather a synergistic gain where joint training produces outcomes far superior to the sum of its parts. For example, in Table I Using IsaacGym as the test environment, we compare policies trained on Genesis, on Sim, and on their combination (Sim+Genesis). Genesis-only training achieves a success rate of 0.043; Sim-only achieves 0.143. Joint training on Sim+Genesis raises the success rate to 0.400 and, across all error metrics, surpasses both single-simulator

baselines as well as their additive baseline.

3) *PolySim enhances in-distribution robustness and precision*: For instance, using **IsaacGym** as the test environment, we compare training solely on **IsaacGym** with a mixed setting (**Gym+Genesis**). Training on Gym alone already achieves a perfect success rate of, but the associated error metrics remain relatively high. In contrast, the Gym+Genesis policy not only preserves the perfect success rate but also substantially reduces these errors respectively, representing the best overall performance in this environment. This result highlights the regularization effect of heterogeneous training: by incorporating diverse dynamics from an additional simulator, the policy avoids overfitting to the simulator inductive bias and instead learns smoother, more generalizable control strategies. Consequently, it achieves higher precision and robustness even when evaluated back on the source domain.

D. Parallel vs. Sequential Multi-Simulator Training

To answer **Q2**, we employ the parallel training framework of PolySim and compare it against policies obtained from sequential training with different simulator permutations, evaluated across five distinct motion tasks. Importantly, all comparisons are conducted under identical RL algorithms and training settings across simulators to ensure fairness. As shown in Fig 4, the results highlight these findings:

1) *Limited improvement from sequential multi-simulator training*: For unseen evaluation simulators, success rates

TABLE II: Iteration times for PolySim compared to the slowest simulator in each recipe. Values in parentheses indicate the additional cost relative to the slowest simulator.

Slowest Simulator	Iteration Times(s)	PolySim Recipe	Iteration Time(s)
IsaacGym	2.1	IsaacGym+IsaacSim	2.3(+0.2)
Genesis	4.7	IsaacGym+Genesis	4.9(+0.2)
Genesis	4.7	IsaacSim+Genesis	4.8(+0.1)
Genesis	4.7	IsaacGym+IsaacSim+Genesis	5.0(+0.3)



Fig. 5: Visualization of sim-to-sim performance in MuJoCo across different training methods. The five panels show the humanoid performing a forward jumping motion under single-simulator, sequential multi-simulator, and PolySim multi-simulator training. Only the PolySim policy trained in parallel across simulators successfully executes the motion in MuJoCo.

show only marginal gains compared to single-simulator training—3.5% and 4.9% improvements on IsaacSim and Genesis, respectively—while performance on IsaacGym and MuJoCo even decreases. For seen evaluation simulators, sequential training across multiple simulators can actually degrade performance. In particular, 3-serial training performs worse than 2-serial by 10.7% and 0.6% on IsaacGym and Genesis, respectively. This anomalous behavior is described as catastrophic forgetting issue and can be observed in Fig 5 also. Our **PolySim** with fully parallel training design significantly improved performance.

E. Experimental efficiency analysis

TABLE II reports iteration times for PolySim compared to the slowest simulator in each recipe. The results show that PolySim’s iteration time remains consistently close to that of the slowest simulator, with only a marginal overhead of +0.1–0.3 s. This demonstrates that PolySim achieves efficient multi-simulator training: it leverages diverse physics engines in parallel while incurring negligible additional computation and communication cost.

F. Sim to real deployment

To answer **Q3**, we deploy policies trained with PolySim—using multiple simulator combinations augmented with parameter-based DR—directly onto the Unitree G1 robot. In all cases, the policies achieve successful zero-shot deployment. Detailed motion demonstrations will be provided in our supplementary video materials.

TABLE III: Motion tracking performance of Kobe under different training environments, evaluated on unseen environments. Here, DR denotes parameter-based Domain Randomization in ASAP [44].

Training Env.	Genesis		MuJoCo	
	Succ \uparrow	$E_{g\text{-}mpjpe}$	Succ \uparrow	$E_{g\text{-}mpjpe}$
IsaacGym _{DR}	1.000	163.135	0.100	295.877
IsaacSim _{DR}	1.000	130.936	0.100	272.610
IsaacSim+IsaacGym	1.000	103.941	0.100	178.190
IsaacSim+IsaacGym+Genesis	-	-	1.000	199.166

VI. CONCLUSIONS

In this work, we proposed **PolySim**, a framework that mitigates simulator inductive bias for robust sim-to-real transfer by training policies across heterogeneous simulators. We provide a theoretical proof that PolySim’s dynamics-level randomization admits a tighter upper bound on this bias, a finding validated by extensive experiments. Our results demonstrate superior zero-shot generalization to unseen simulators (e.g., MuJoCo) and enhanced in-domain performance, culminating in the successful zero-shot deployment of a policy onto the Unitree G1 humanoid robot. This work establishes **PolySim** as a scalable pathway for transferring policies to real-world robots; future work will expand it with more simulators and diverse tasks like manipulation.

REFERENCES

- [1] M. Yuan, T. Yu, W. Ge, X. Yao, D. Li, H. Wang, J. Chen, X. Jin, B. Li, H. Chen *et al.*, “Behavior foundation model: Towards next-

- generation whole-body control system of humanoid robots,” *arXiv preprint arXiv:2506.20487*, 2025.
- [2] X. Chen, J. Hu, C. Jin, L. Li, and L. Wang, “Understanding domain randomization for sim-to-real transfer,” in *International Conference on Learning Representations*.
- [3] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, “Domain randomization for transferring deep neural networks from simulation to the real world,” in *International conference on intelligent robots and systems (IROS)*. IEEE, 2017, pp. 23–30.
- [4] H. Geng, F. Wang, S. Wei, Y. Li, B. Wang, B. An, C. T. Cheng, H. Lou, P. Li, Y.-J. Wang, Y. Liang, D. Goetting, C. Xu, H. Chen, Y. Qian, Y. Geng, J. Mao, W. Wan, M. Zhang, J. Lyu, S. Zhao, J. Zhang, J. Zhang, C. Zhao, H. Lu, Y. Ding, R. Gong, Y. Wang, Y. Kuang, R. Wu, B. Jia, C. Sferrazza, H. Dong, S. Huang, Y. Wang, J. Malik, and P. Abbeel, “Roboverse: Towards a unified platform, dataset and benchmark for scalable and generalizable robot learning,” *arXiv preprint arXiv:2504.18904*, 2025. [Online]. Available: <https://arxiv.org/abs/2504.18904>
- [5] LeCAR-Lab, “Humanoidverse: A multi-simulator framework for humanoid robot sim-to-real learning,” <https://github.com/LeCAR-Lab/HumanoidVerse>, 2025, GitHub repository.
- [6] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” *arXiv preprint arXiv:1707.06347*, 2017.
- [7] V. Makovychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa *et al.*, “Isaac gym: High performance gpu-based physics simulation for robot learning,” *arXiv preprint arXiv:2108.10470*, 2021.
- [8] E. Todorov, T. Erez, and Y. Tassa, “Mujoco: A physics engine for model-based control,” in *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2012, pp. 5026–5033.
- [9] M. Mittal, C. Yu, Q. Yu, J. Liu, N. Rudin, D. Hoeller, J. L. Yuan, R. Singh, Y. Guo, H. Mazhar *et al.*, “Orbit: A unified simulation framework for interactive robot learning environments,” *IEEE Robotics and Automation Letters*, vol. 8, no. 6, pp. 3740–3747, 2023.
- [10] T. Mu, Z. Ling, F. Xiang, D. Yang, X. Li, S. Tao, Z. Huang, Z. Jia, and H. Su, “Maniskill: Generalizable manipulation skill benchmark with large-scale demonstrations,” *arXiv preprint arXiv:2107.14483*, 2021.
- [11] T. Li, H. Geyer, C. G. Atkeson, and A. Rai, “Using deep reinforcement learning to learn high-level policies on the atrias biped,” in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 263–269.
- [12] 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)*. IEEE, 2021, pp. 2811–2817.
- [13] Q. Liao, B. Zhang, X. Huang, X. Huang, Z. Li, and K. Sreenath, “Berkeley humanoid: A research platform for learning-based control,” in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 2897–2904.
- [14] H. Wang, Z. Wang, J. Ren, Q. Ben, T. Huang, W. Zhang, and J. Pang, “Beamndojo: Learning agile humanoid locomotion on sparse footholds,” *arXiv preprint arXiv:2502.10363*, 2025.
- [15] X. Gu, Y.-J. Wang, and J. Chen, “Humanoid-gym: Reinforcement learning for humanoid robot with zero-shot sim2real transfer,” *arXiv preprint arXiv:2404.05695*, 2024.
- [16] X. He, R. Dong, Z. Chen, and S. Gupta, “Learning getting-up policies for real-world humanoid robots,” *arXiv preprint arXiv:2502.12152*, 2025.
- [17] Z. Li, X. B. Peng, P. Abbeel, S. Levine, G. Berseth, and K. Sreenath, “Robust and versatile bipedal jumping control through reinforcement learning,” *arXiv preprint arXiv:2302.09450*, 2023.
- [18] J. Long, J. Ren, M. Shi, Z. Wang, T. Huang, P. Luo, and J. Pang, “Learning humanoid locomotion with perceptive internal model,” in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 9997–10 003.
- [19] Z. Zhuang, S. Yao, and H. Zhao, “Humanoid parkour learning,” *arXiv preprint arXiv:2406.10759*, 2024.
- [20] C. Zhang, W. Xiao, T. He, and G. Shi, “Wococo: Learning whole-body humanoid control with sequential contacts,” *arXiv preprint arXiv:2406.06005*, 2024.
- [21] M. Ji, X. Peng, F. Liu, J. Li, G. Yang, X. Cheng, and X. Wang, “Exbody2: Advanced expressive humanoid whole-body control,” *arXiv preprint arXiv:2412.13196*, 2024.
- [22] X. Cheng, Y. Ji, J. Chen, R. Yang, G. Yang, and X. Wang, “Expressive whole-body control for humanoid robots,” *arXiv preprint arXiv:2402.16796*, 2024.
- [23] C. Tessler, Y. Guo, O. Nabati, G. Chechik, and X. B. Peng, “Masked-mimic: Unified physics-based character control through masked motion inpainting,” *ACM Transactions on Graphics (TOG)*, vol. 43, no. 6, pp. 1–21, 2024.
- [24] X. B. Peng, P. Abbeel, S. Levine, and M. Van de Panne, “Deepmimic: Example-guided deep reinforcement learning of physics-based character skills,” *ACM Transactions On Graphics (TOG)*, vol. 37, no. 4, pp. 1–14, 2018.
- [25] S. Xu, H. Y. Ling, Y.-X. Wang, and L.-Y. Gui, “Intermimic: Towards universal whole-body control for physics-based human-object interactions,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 12 266–12 277.
- [26] J. Tan, T. Zhang, E. Coumans, A. Iscen, Y. Bai, D. Hafner, S. Bohez, and V. Vanhoucke, “Sim-to-real: Learning agile locomotion for quadruped robots,” *arXiv preprint arXiv:1804.10332*, 2018.
- [27] B. Mehta, M. Diaz, F. Golemo, C. J. Pal, and L. Paull, “Active domain randomization,” in *Conference on Robot Learning*. PMLR, 2020, pp. 1162–1176.
- [28] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, “Domain generalization: A survey,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 4, pp. 4396–4415, 2022.
- [29] F. Kozin and H. Natke, “System identification techniques,” *Structural safety*, vol. 3, no. 3-4, pp. 269–316, 1986.
- [30] K. J. Åström and P. Eykhoff, “System identification—a survey,” *Automatica*, vol. 7, no. 2, pp. 123–162, 1971.
- [31] L. Ljung, “Perspectives on system identification,” *Annual Reviews in Control*, vol. 34, no. 1, pp. 1–12, 2010.
- [32] J. Hwangbo, J. Lee, A. Dosovitskiy, D. Bellicoso, V. Tsounis, V. Koltun, and M. Hutter, “Learning agile and dynamic motor skills for legged robots,” *Science Robotics*, vol. 4, no. 26, p. eaau5872, 2019.
- [33] P. K. Khosla and T. Kanade, “Parameter identification of robot dynamics,” in *1985 24th IEEE conference on decision and control*. IEEE, 1985, pp. 1754–1760.
- [34] W. Yu, V. C. Kumar, G. Turk, and C. K. Liu, “Sim-to-real transfer for biped locomotion,” in *2019 ieee/rsj international conference on intelligent robots and systems (iros)*. IEEE, 2019, pp. 3503–3510.
- [35] Y. Du, O. Watkins, T. Darrell, P. Abbeel, and D. Pathak, “Auto-tuned sim-to-real transfer,” in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 1290–1296.
- [36] Y. Chebotar, A. Handa, V. Makovychuk, M. Macklin, J. Issac, N. Ratliff, and D. Fox, “Closing the sim-to-real loop: Adapting simulation randomization with real world experience,” in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8973–8979.
- [37] W. Yu, C. K. Liu, and G. Turk, “Policy transfer with strategy optimization,” *arXiv preprint arXiv:1810.05751*, 2018.
- [38] X. B. Peng, E. Coumans, T. Zhang, T.-W. Lee, J. Tan, and S. Levine, “Learning agile robotic locomotion skills by imitating animals,” *arXiv preprint arXiv:2004.00784*, 2020.
- [39] K.-H. Lee, O. Nachum, T. Zhang, S. Guadarrama, J. Tan, and W. Yu, “Pi-ars: Accelerating evolution-learned visual-locomotion with predictive information representations,” in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 1447–1454.
- [40] W. Yu, J. Tan, Y. Bai, E. Coumans, and S. Ha, “Learning fast adaptation with meta strategy optimization,” *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2950–2957, 2020.
- [41] W. Yu, J. Tan, C. K. Liu, and G. Turk, “Preparing for the unknown: Learning a universal policy with online system identification,” *arXiv preprint arXiv:1702.02453*, 2017.
- [42] A. Kumar, Z. Fu, D. Pathak, and J. Malik, “Rma: Rapid motor adaptation for legged robots,” *arXiv preprint arXiv:2107.04034*, 2021.
- [43] A. Kumar, Z. Li, J. Zeng, D. Pathak, K. Sreenath, and J. Malik, “Adapting rapid motor adaptation for bipedal robots,” in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 1161–1168.
- [44] T. He, J. Gao, W. Xiao, Y. Zhang, Z. Wang, J. Wang, Z. Luo, G. He, N. Sobanbab, C. Pan *et al.*, “Asap: Aligning simulation and real-world physics for learning agile humanoid whole-body skills,” *arXiv preprint arXiv:2502.01143*, 2025.