MSK-Bench: Benchmarking Full-Body Musculoskeletal Motor Control Across Tasks, Learning Paradigms, and Physiological Metrics

Author Names Omitted for Anonymous Review. Paper-ID [1]

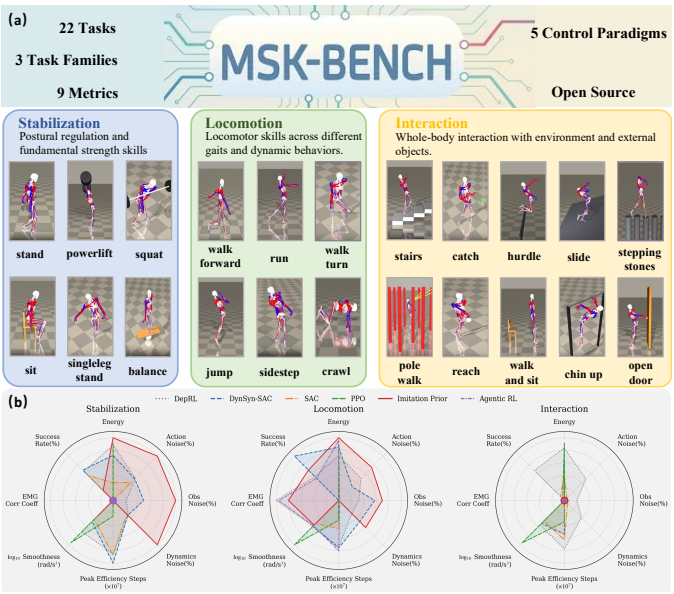

Fig. 1: **Overview of the Musculoskeletal Benchmark (MSK-Bench).** (a) **Benchmark Components:** An open-source suite featuring **22** tasks across stabilization, locomotion, and interaction, evaluated under **5** control paradigms and **9** metrics. (b) **Method Evaluation:** Radar charts compare control methods across the **3** task families based on **8** performance and robustness axes. Each value represents the average performance across all tasks within the respective category. Note that due to the scarcity of human reference data, the EMG correlation coefficient is only evaluated on the walk task, and the remaining cumulative return metric is detailed in the Fig. 2.

***Abstract*—Musculoskeletal (MSK) humanoids provide a physiologically grounded embodiment for studying full-body motor control, but their high-dimensional muscle actuation, delayed activation dynamics, and redundant muscle-tendon structures make learning substantially harder than torque-driven humanoid control. Existing MSK benchmarks remain fragmented across gait, prosthetics, dexterous hands, or challenge-specific tracks, leaving full-body muscle-actuated control insufficiently evaluated under standardized tasks, methods, and metrics. We introduce MSK-Bench, a benchmark of 22 full-body motor-control tasks organized into three progressively challenging categories: postural stabilization, common locomotor behaviors, and contact-rich environmental interaction. Under unified task protocols and robustness perturbations, MSK-Bench evaluates 5 representative control paradigms, including reward-based RL, agentic reward tuning, latent-action RL, imitation-prior control, and residual adaptation over imitation priors. Beyond task success and reward, MSK-Bench further reports robustness analysis and physiology-oriented diagnostics, including activation cost, joint smoothness, and walking EMG-envelope similarity. Our empirical study shows that embodiment-aware exploration improves task coverage in high-dimensional muscle spaces, imitation priors enhance reference-compatible stabilization and locomotion but degrade under contact-rich terrain mismatch, and residual adaptation can recover successful behaviors when fixed references fail. MSK-Bench provides a task–method–metric testbed for full-body muscle-actuated humanoid control.**

I. Introduction

Human motion emerges from the coordinated control of hundreds of muscles acting through a highly redundant and biomechanically constrained musculoskeletal system [112, 58]. Unlike torque-driven humanoids, where actions directly specify joint torques, biological movement is generated through muscle excitation, activation-contraction dynamics and tendon transmission [22, 113]. This physiological hierarchy introduces substantial control challenges including delayed actuation, muscle redundancy, and complex nonlinear dynamics [95]. Consequently, musculoskeletal (MSK) models have become an increasingly important platform for studying physically grounded control [80, 11, 19].

Despite this progress, our understanding of how to learn effective control policies for large-scale musculoskeletal systems remains limited. Recent torque-driven humanoid benchmarks have substantially expanded task diversity for whole-body locomotion, loco-manipulation, and scene interaction, but they typically rely on rigid-body torque actuation (Tab. I). Conversely, existing MSK benchmarks provide physiologically grounded actuation, but they remain largely fragmented across specific domains, including gait optimization [85, 44], prosthetics [46], athletic locomotion [13, 99], and dexterous hand control [12, 107]. While these benchmarks have significantly advanced individual subfields, they are still limited in at least one of the dimensions needed for general-purpose full-body muscle-actuated control: broad whole-body task diversity, contact-rich terrain and object interaction, or standardized evaluation across multiple control paradigms. As a result, it remains unclear which control paradigms are most effective for full-body musculoskeletal agents and under what physical conditions they succeed or fail.

A second limitation of existing research lies in how musculoskeletal control is evaluated. Most existing benchmarks primarily focus on task-level metrics such as reward, success

rate, or locomotion speed [26]. However, unlike conventional humanoid control, musculoskeletal systems provide access to physiologically meaningful internal variables, including muscle activations, muscle synergies, and energy expenditure [20, 78]. Multiple muscle patterns can often generate similar body motions, implying that successful task completion alone is insufficient for evaluating musculoskeletal control [95, 50]. A policy may achieve the correct movement while relying on unrealistic muscle coordination strategies. Therefore, evaluating whether a policy behaves in a biomechanically plausible manner is equally important as evaluating whether it completes a task successfully [60, 83].

To address these challenges, we propose **MSK-Bench**, a comprehensive full-body musculoskeletal benchmark designed to resolve current evaluation deficiencies: the lack of humanoid-level task diversity, the scarcity of rich physical interactions, and the absence of standardized physiological metrics (Tab. I provides a detailed comparison) [81, 47, 56, 114, 37, 83, 6, 12, 51, 117, 104, 7, 14, 97]. Specifically, MSK-Bench encompasses **22** tasks organized into three progressively challenging categories: stabilization, locomotion, and complex interactions (e.g., load transport and whole-body coordination), as illustrated in Fig. 1a. Moving beyond mere task success, we establish a physiology-aware evaluation protocol that assesses biomechanical fidelity through standardized metrics, including muscle activation cost, joint smoothness, robustness, and walking EMG-envelope similarity. We evaluate **five** representative control paradigms under this unified protocol to provide a comprehensive multi-dimensional analysis of their control capabilities (Fig. 1b). This evaluation establishes a strong baseline, demonstrating that residual reinforcement learning effectively combines motor priors with task adaptation to offer a robust foundation for future research.

Through **MSK-Bench**, we identify three empirical findings about full-body muscle-actuated control. **First**, broad task coverage in highly overactuated MSK systems benefits from embodiment-aware exploration and action-structured representations, while unconstrained reward retuning and overly aggressive action compression do not reliably solve the exploration challenge. **Second**, motion priors provide strong robustness for reference-compatible stabilization and locomotion skills, but become brittle under contact-rich terrain mismatch where fixed reference timing and foot-placement patterns no longer align with the physical environment. **Third**, physiology-oriented diagnostics reveal that task demands, policy class, and anatomical resolution all shape internal muscle recruitment patterns, highlighting why task success alone is insufficient for evaluating MSK control. These findings further motivate residual adaptation as a strong baseline that preserves the stabilizing benefits of imitation priors while learning task-driven corrections for contact and muscle-use limitations.

II. Related Work

Musculoskeletal Models and Control. Musculoskeletal (MSK) agents differ from torque-driven humanoids: muscle excitations are transformed by activation-contraction dynamics into muscle forces and then mapped to joint torques through tendon paths and moment arms, introducing redundancy, delay, nonlinear muscle-tendon dynamics, and biomechanical constraints [22, 79, 80]. Classical platforms established workflows for anatomically grounded model construction, scaling, and motion-data analysis [22, 79, 80, 19], while MyoSim and MyoSuite brought physiologically realistic MSK models into MuJoCo-style contact-rich RL settings [93, 10, 98, 11]. Recent work further advances MSK control through model-based muscle controllers, hierarchical planning, markerless-capture imitation, and physical musculoskeletal humanoids [27, 102, 17, 43, 107], while full-body systems have scaled toward large-scale motion imitation, richer anatomy, and whole-body behavior emulation [51, 117, 104]. These advances motivate a systematic evaluation of full-body MSK control.

Benchmarks for Embodied Motor Control. Robot-learning benchmarks have standardized progress in continuous control, offline RL [90, 29], and manipulation [110, 39, 59, 40, 54, 116, 48, 32, 89, 64, 65, 62, 96, 53]. Recent humanoid benchmarks further move from isolated locomotion or manipulation toward whole-body interaction [81, 47, 56, 114, 37]. However, these benchmarks primarily target torque-driven rigid-body robots rather than physiological muscle actuation. For the MSK benchmark, early OpenSim-based challenges focused on gait, running, prosthetics, and neuromechanical locomotion control [44, 45, 46, 85, 2]. MyoSuite [11] and MyoChallenge [14, 97, 13, 99] introduce contact-rich MSK tasks for physiologically realistic models, spanning hand manipulation, locomotion, and athletic intelligence tracks. Our benchmark organizes full-body MSK tasks into a graded spectrum from postural regulation to common locomotor behaviors and contact-rich environmental interaction. It evaluates multiple control paradigms under a task-method-metric matrix, where metrics go beyond task success to include EMG-envelope activation [51, 7, 83] for assessing biomechanically plausible muscle coordination.

Learning Paradigms for Musculoskeletal Control. Reward-based RL provides a general route to motor control [76, 33], but the high dimensionality, redundancy, and delay of muscle actuation make direct exploration inefficient in MSK agents, motivating MSK-specific exploration, action-structure, and weak-prior methods [77, 36, 103, 6, 12]. Agentic reward generation further reduces manual reward engineering in robotics [111, 106, 57, 100], yet its reliability under physiologically constrained MSK dynamics remains unclear. A complementary family relies on motion references and motor priors: references can be reconstructed from human motion or generated by large motion models [94, 31, 75, 105, 101], while imitation and prior-based controllers have enabled both physics-based character control and muscle-actuated motion imitation [51, 83, 69, 71, 72, 91, 63]. However, reference tracking alone may fail on contact-rich or out-of-distribution tasks, motivating residual adaptation that learns corrective actions on top of existing controllers or skill policies [41, 82, 3, 74, 52, 108, 4, 115, 87, 109, 35]. Our benchmark jointly evaluates reward-based RL, agentic reward tuning, latent-

TABLE I: Comparison of **MSK-Bench** with existing **torque-driven humanoid** (top) and **musculoskeletal** (bottom) benchmarks. To our knowledge, MSK-Bench is the first to jointly support high-dimensional muscle actuation, whole-body modeling, complex terrain, contact-rich interaction, and multi-paradigm physiological evaluation. ✓/●/✗ denote full/partial/no support. **Columns: Actuation (Dim.)**: actuation type and action dimension; **Full Body**: whole-body embodiment rather than isolated limbs or hands; **3D Scene/Terrain**: uneven terrain or 3D obstacles; **Contact Rich**: object interaction or multi-limb environmental contact; **Task Hierarchy**: structured task families from stabilization to locomotion and interaction; **#Tasks/Data Scale**: number of environments, tasks, or motion clips; **Physio. Eval. Metrics**: physiological diagnostics beyond reward, such as activation cost, smoothness, and EMG similarity; **Evaluated Control Families**: officially benchmarked learning/control families.

Benchmark / Framework	Embodiment		Environment		Task Design		Evaluation		Main Focus
	Actuation (Dim.)	Full Body	3D Scene / Terrain	Contact Rich	Task Hierarchy	#Tasks / Data Scale	Physio. Eval. Metrics	Evaluated Families	
Torque-Driven Humanoid Benchmarks									
OpenAI Gym Humanoid [10]	Torque (∼17)	✓	✗	✗	✗	1 task	✗	1	Locomotion
HumanoidBench [81]	Torque (19/61)	✓	✓	✓	✓	27 tasks	✗	1	Loco-manipulation
SkillBench / SkillBlender [47]	Torque (19)	✓	●	✓	✓	4 skills / 8 tasks	✗	1	Skill Blending
Mimicking-Bench [56]	Torque (19)	✓	✓	✓	✗	6 tasks	✗	1	Scene Interaction
OmniH2O [37]	Torque (19)	✓	●	✓	●	6 tasks	✗	1	[Real-World] Teleoperation
LHM-Humanoid [114]	Torque (19)	✓	✓	✓	✓	350 tasks	✗	1	Rearrangement
Musculoskeletal Benchmarks									
LocoMuJoCo [2]	Mixed	✓	✗	✗	●	12 envs / 27 tasks	✗	2	Imitation (Loco)
MyoSuite [11]	Muscle (≤80)	✗	✗	✓	●	204 tasks	✗	1	Dexterity & Agility
MyoDex [12]	Muscle (39)	✗	✗	✓	✗	14 train / 34 eval tasks	✗	1	Hand Manipulation
MS-Human-700 / MsGym [117]	Muscle (700)	✓	✗	●	✗	3 tasks	✓	1	MSK Locomotion
MyoChallenge 2024 [99]	Muscle (≤80)	●	✓	✓	●	2 tracks	✗	1	Prosthetics
MyoChallenge 2025 [97]	Muscle (≤80)	●	●	✓	●	2 tasks / 4 tracks	✗	1	Athletic Control
Musculoskeletal-Bench (Ours)	**Muscle (416/700)**	✓	✓	✓	✓	**22 tasks**	✓	**5**	Full-body Terrain MSK Control

action RL, imitation-prior control, and residual adaptation under the same full-body MSK task suite.

III. MSK-Bench

As shown in Fig. 1a, MSK-Bench is designed around three foundational functional regimes of gross full-body motor control: maintaining postural stability, generating self-propelled movement, and coordinating physical interaction with external terrain or objects [67]. These regimes progressively broaden the control demands from internal body stabilization to dynamic whole-body locomotion and contact-rich environmental interaction. Accordingly, our 22 tasks are organized into three task families (details in App. A): 1) **Stabilization** evaluates whether an MSK controller can maintain static or quasi-static whole-body posture under body weight and external loading, covering tasks such as standing, single-leg standing, squatting, sitting, balancing, and powerlifting [42, 9, 16]. These tasks emphasize postural regulation, core control, and trunk stabilizer engagement [9, 86]. 2) **Locomotion** evaluates self-propelled movement skills, including walking, running, jumping, crawling, sidestepping, and turning [49, 66, 112]. These tasks require coordinated lower-limb recruitment, dynamic balance, gait-phase coordination, and impact handling across cyclic and non-cyclic movements [55, 34]. 3) **Interaction** evaluates whole-body motor control under external contact, terrain variation, and object interaction, including stairs, hurdles, stepping stones, sliding terrain, reaching, catching, chin-up, door opening, pole walking, and walk-and-sit tasks [11, 81]. These tasks require the controller to coordinate lower-body support, trunk stabilization, upper-limb contact, and grip or pulling forces while adapting to physical constraints imposed by the environment [61, 25, 84]. Together, these task families provide a structured benchmark for evaluating whether muscle-actuated agents can balance, move, and physically interact while producing biomechanically plausible muscle activations across diverse full-body movement regimes [81, 11, 95].

Experimental Setup and Evaluation Protocol. All tasks share a common muscle-action interface, simulator configuration, and base proprioceptive observation schema, with task-specific goal, terrain, or object observations added when required (details in App. A). Performance is measured over multiple test episodes using held-out random seeds. To comprehensively assess algorithmic capabilities, we evaluate baselines using a spectrum of task-level and physiology-oriented metrics (details in App. B). Task-level performance is quantified via **Success Rate** (fulfilling task criteria without constraint violations), **Cumulative Reward** (measuring overall task performance and efficiency via step-wise reward aggregation), and **Peak Efficiency Steps** (evaluating sample efficiency for peak skill acquisition). We also assess zero-shot **Robustness** under Sim2Real-style perturbations, including **action noise**, **observation noise**, and **dynamics randomization**. To ensure physiological fidelity, we report physiology-oriented metrics:

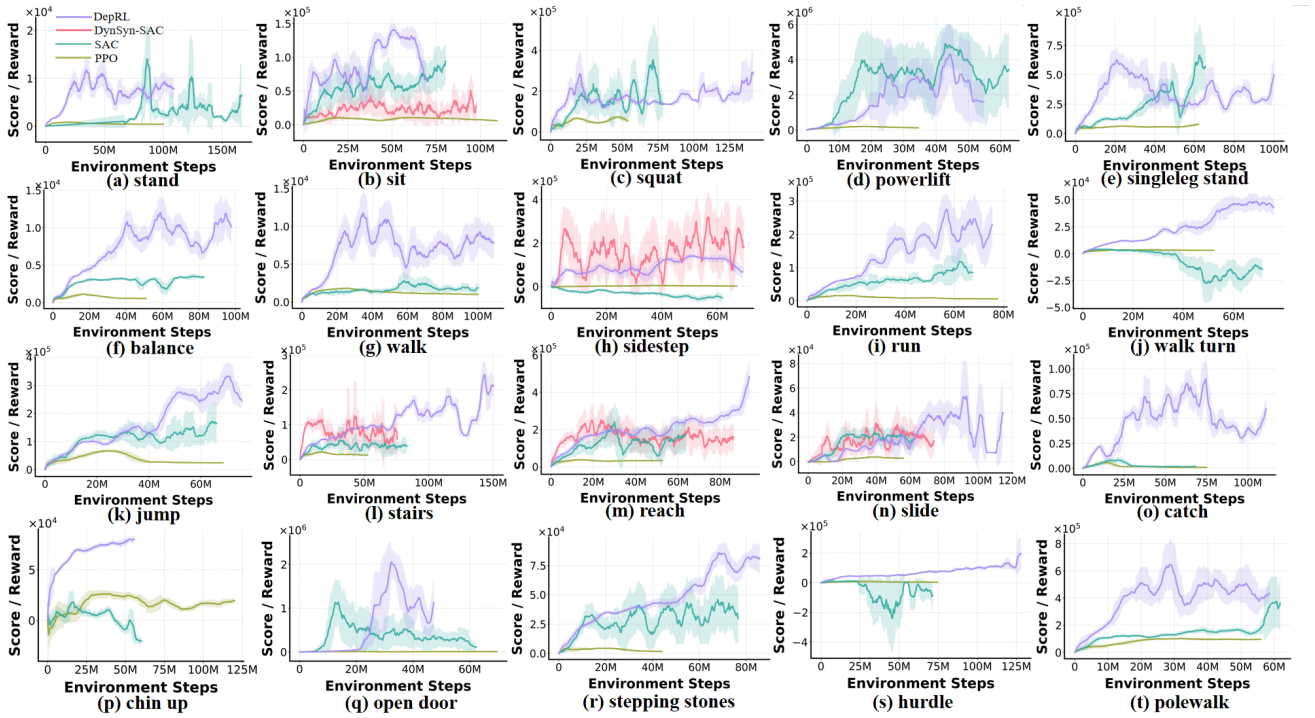

Fig. 2: **Learning curves across 20 of the 22 benchmark tasks for DepRL, DynSyn-SAC, SAC, and PPO.** DepRL achieves the broadest task coverage under the fixed training budget, while PPO and SAC often fail to obtain stable high-return behaviors on dynamic locomotion and contact-rich tasks. DynSyn-SAC performs competitively on selected action-structured tasks such as sidestepping. Shaded regions indicate variation across seeds.

Muscle Activation Energy as a proxy for physiological efficiency to filter out excessive co-contraction, **Joint Smoothness** (quantified by finite-difference jerk) for coordinated movement, and **EMG Similarity** to measure the waveform correlation between simulated muscle activations and human EMG dataset recordings [7].

Baselines. To provide a comprehensive evaluation of algorithmic capabilities on MSK-Bench, we benchmark a diverse spectrum of control strategies: 1) **Reward-based RL**, comprising standard and exploration-enhanced methods (details in App. C and D) such as PPO [76], SAC [33], DepRL [77], and DynSyn-SAC [36]; 2) **Agentic reward tuning**, which uses DepRL [77] as the base learner and employs DeepSeek-V4-Flash [21] to iteratively modify reward weights based on observed failure modes; 3) **Latent-action RL**, which compresses the original high-dimensional muscle excitation space into lower-dimensional (128 to 16) control representations before training; 4) **Imitation-prior control**, which uses a pretrained muscle-actuated policy from MuscleMimic [51] learned from retargeted reference motions with joint-space and mimic-site tracking rewards; and 5) **Residual adaptation over imitation priors**, which learns task-specific corrective muscle actions superimposed on the pretrained imitation baseline.

Discussion. IWe use MSK-Bench to answer three questions. First, *how should high-dimensional muscle action spaces be explored?* Broad task coverage benefits from embodiment-aware exploration and action-structured representations, whereas unconstrained reward retuning and overly aggressive action compression are not sufficient on their own **(T1)**. Second, *where do motion priors succeed or fail?* They improve robustness on reference-compatible stabilization and locomotion, but become brittle under terrain-induced contact mismatch **(T2)**. Third, *what is missed by task success alone?* Physiology-oriented diagnostics reveal task-, policy-, load-, and anatomy-dependent differences in movement smoothness, EMG-envelope alignment, and muscle recruitment **(T3)**.

Takeaway 1 (T1): Embodiment-Aware Exploration and Action Structure Improve Broad Task Coverage, While Naive Reward Retuning and Aggressive Compression Remain Unstable. Musculoskeletal agents are highly overactuated: many muscles act on fewer mechanical degrees of freedom, agonist–antagonist pairs can cancel each other, and short uncoordinated activations may be filtered by muscle activation dynamics before producing meaningful joint-space motion [77, 36]. As a result, exploration in the raw muscle excitation space is difficult. Under our fixed training budget, standard unstructured reward-based RL baselines often fail to discover coordinated multi-muscle recruitment across the full task suite, whereas DepRL achieves broader task coverage by exploiting embodiment-aware sensor–action couplings (Fig. 2). DynSyn-SAC further illustrates that action-structured representations can be beneficial in tasks such as sidestepping, where coordinated lateral muscle recruitment is critical (see all metrics in Tab. XI, Tab. XII and Tab. XIII).

We next examine two common interventions for improving learning in high-dimensional action spaces: **automated reward retuning** (App. F) and **action-space compression** (App. G). In our LLM-guided reward-tuning experiment, high-

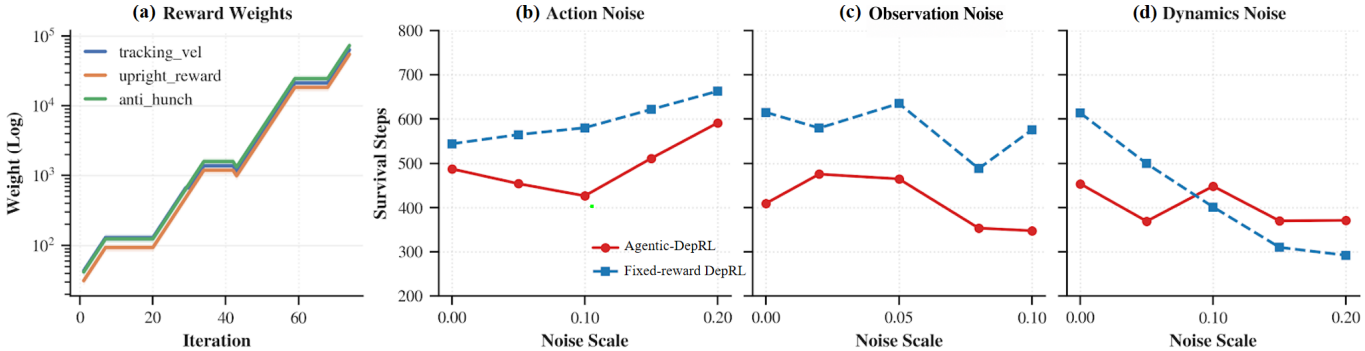

Fig. 3: **Online LLM-guided reward-weight tuning does not consistently improve robustness over fixed-reward DepRL.** (a) Evolution of reward weights updated online during training; the reward terms are fixed and only scalar weights are modified. (b–d) Survival steps under **action**, **observation**, and **dynamics** perturbations for Agentic-DepRL and fixed-reward DepRL. Both variants use the same DepRL [77] base algorithm, training budget, and policy architecture. Survival is capped at 1000 steps; each point averages 50 held-out evaluation seeds. More details are given in App. B and App. F.

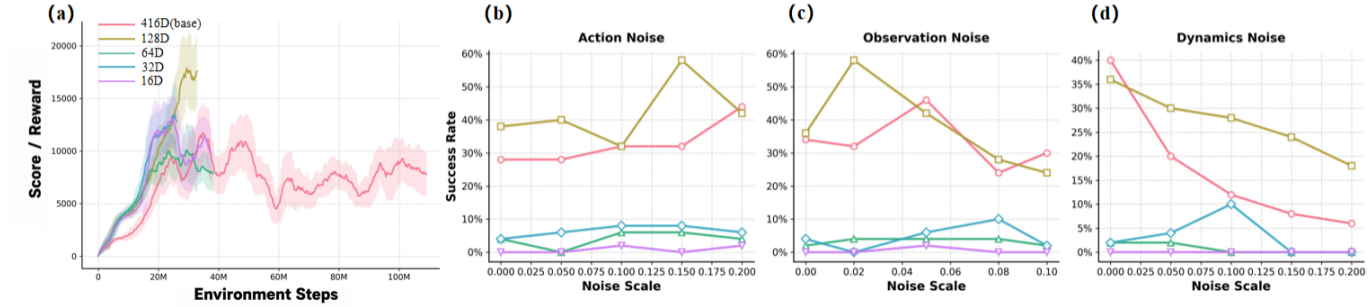

Fig. 4: **Moderate action-space compression helps, but aggressive compression loses expressivity.** Reducing the 416-dimensional muscle-control space to 128 dimensions improves reward learning and several robustness settings, whereas stronger compression to 64, 32, or 16 dimensions substantially degrades performance (details in App. G).

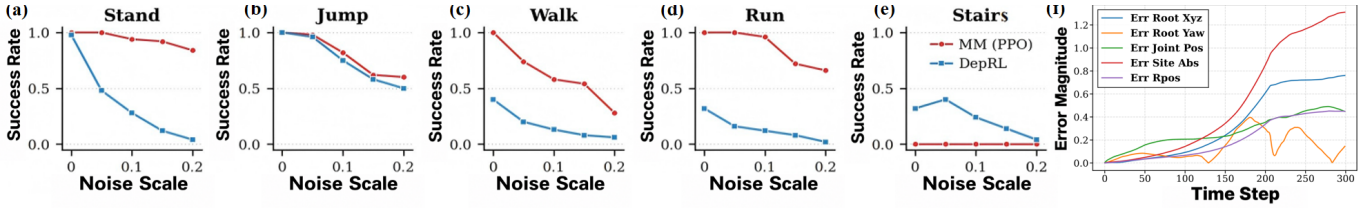

Fig. 5: **Robustness comparison and tracking error analysis between Imitation prior and DepRL.** (a–e) Success rates of MuscleMimic/MM (PPO) [51] and DepRL [77] under increasing dynamic perturbation scales on (a) Stand, (b) Jump, (c) Walk, (d) Run, and (e) Stairs. MM (PPO) shows stronger robustness on Stand, Jump, Walk, and Run, whereas DepRL performs better on Stairs. (f) Tracking error magnitude of MuscleMimic on the **Stairs** task over time. The accumulated tracking errors reveal the limitation of imitation-based policies when handling complex terrain.

frequency and unconstrained reward-weight updates produce a strongly non-stationary objective. Fig. 3a shows rapid reward-weight inflation over iterations, while Fig. 3b-d shows that this retuning does not yield consistent robustness gains under action, observation, or dynamics perturbations. This suggests that automated reward design for MSK control requires additional constraints, such as bounded updates, validation-based selection, or lower-frequency retuning, rather than unconstrained online weight adjustment. Action-space compression yields a more nuanced outcome. As shown in Fig. 4, reducing the 416-dimensional muscle-control space to 128 dimensions improves reward learning and robustness in several perturbation settings, indicating that moderate low-dimensional structure can reduce the exploration burden. However, more aggressive reductions to 64, 32, or 16 dimensions substantially degrade performance, suggesting that full-body MSK control still requires sufficient action bandwidth to coordinate distributed muscles across posture, locomotion, and contact-rich behavior. Together, these results indicate that broad MSK control benefits from embodiment-aware or action-structured priors, but cannot be reliably solved by unconstrained reward retuning or overly aggressive action compression alone.

Takeaway 2 (T2): Motion Priors Improve Flat-Ground Robustness but Are Vulnerable to Contact-Induced Track-

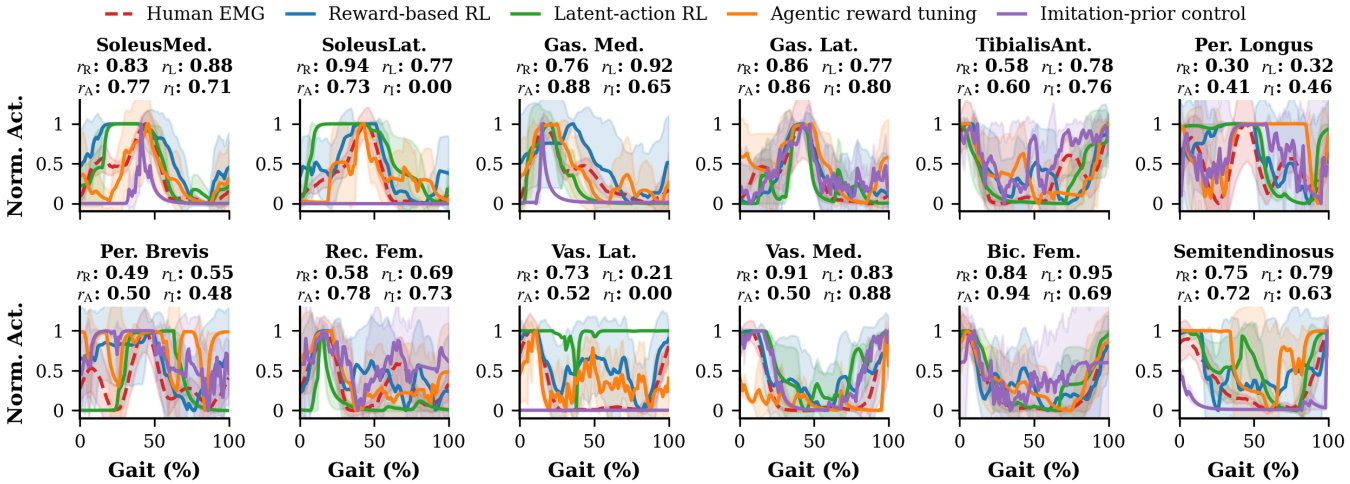

Fig. 6: **Walking EMG-envelope similarity across policies.** Simulated muscle activation envelopes are compared with normalized human walking EMG data [7]. Pearson correlations (r_R, r_L, r_A, r_I) measure gait-cycle waveform similarity for reward-based RL, latent-action RL, agentic reward tuning, and imitation-prior control, respectively (details in App. H).

ing Mismatch. Motion imitation provides a complementary route to MSK control by reducing the exploration burden and supplying phase-structured postural and gait priors. As demonstrated in Fig. 5a-e, reference-motion pretraining (MuscleMimic [51]) provides an inductive bias that reward-driven exploration (DepRL [77]) does not fully replicate across general stabilization and flat-ground locomotion tasks.

However, contact-rich environment interaction tasks, such as Stairs, expose limitations in imitation priors. As shown in Fig. 5f, the imitation policy accumulates tracking errors (particularly in absolute site coordinates and joint positions) over time. While the policy can compensate for this drift on flat ground, the accumulated errors lead to misplaced foot contacts or off-timing force generation on complex terrains. These environments require precise contact strategies that a fixed reference-motion distribution struggles to execute once tracking errors compound. In these scenarios, the reward-driven exploration of DepRL achieves higher robustness during physical interactions.

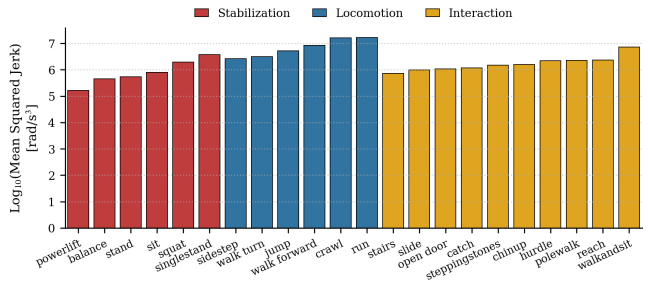

Fig. 7: **Task-wise joint smoothness across benchmark tasks.** Bars show the $\log_{10}$ mean squared joint jerk computed from finite differences of joint velocities. Stabilization tasks exhibit the lowest jerk, locomotion tasks show the highest jerk due to repeated rhythmic impacts, and contact-rich interaction tasks fall in an intermediate range. This diagnostic characterizes task-dependent movement smoothness and impact profile.

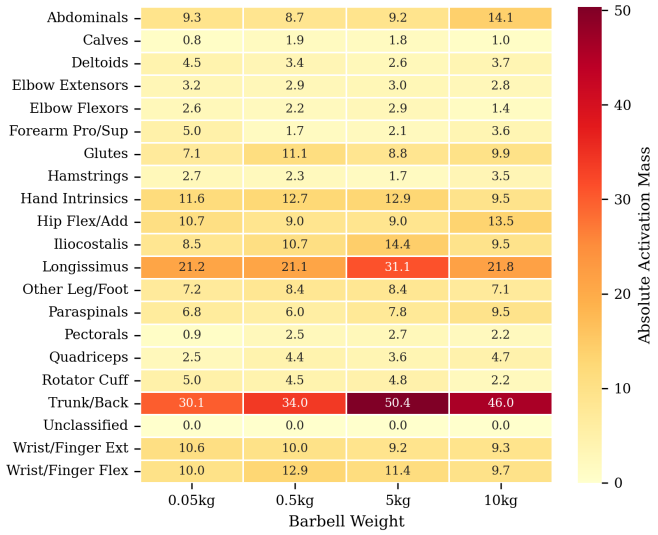

Fig. 8: **Load-dependent muscle-family activation during the powerlift task.** Heatmap values show absolute activation mass accumulated by muscle family under increasing barbell weights. Heavier loads induce stronger recruitment of trunk/back and support-related muscle groups, illustrating how external loading changes the policy's internal muscle-use strategy. (Muscle classification details in App. I)

Takeaway 3 (T3): Physiology-Oriented Diagnostics Reveal Task- and Anatomy-Dependent Muscle Recruitment. Our analysis indicates a connection between the external physical demands of a task and internal biomechanical fidelity. At the macro-kinematic level, analyzing mean squared jerk (Fig. 7) demonstrates that movement smoothness is largely dictated by a task's physical impact profile rather than abstract complexity. Stabilization tasks predictably occupy the lowest jerk spectrum. Conversely, locomotion tasks consistently exhibit the highest jerk due to continuous rhythmic weight transfers and repeated ground reaction forces [24, 112]. Complex environmental interaction tasks exhibit moderate jerk levels,

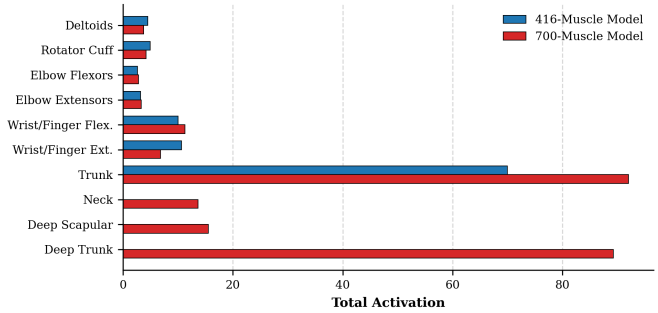

Fig. 9: **Effect of anatomical resolution on powerlift muscle recruitment.** Total activation mass is compared between the 416-muscle model over shared muscle families and the 700-muscle model, which additionally includes neck, deep-trunk, and deep-scapular muscle groups. The result shows that increasing anatomical resolution changes the available recruitment pathways and enables the policy to recruit additional axial stabilizers during load support.

remaining strictly below those of dynamic locomotion. This suggests that the controller utilizes anticipatory postural adjustments (APAs) [58, 8], isolating high-frequency actuations to specific manipulating limbs while stabilizing the rest of the body to minimize kinetic perturbation.

To execute these postural adjustments and manage physical impacts, the controller optimizes internal force generation. Our results indicate that human-like muscle coordination is an optimal biomechanical solution to these demands [18]. Due to musculoskeletal redundancy, multiple sub-optimal strategies can generate identical joint kinematics [83, 51]. Consequently, relying solely on an imitation prior yields accurate kinematics but often fails to recover human-like internal muscle recruitment (Fig. 6). In contrast, optimizing for task completion (e.g., DepRL) guides the policy toward physiologically aligned solutions (e.g., accurate SoleusLat. and Vas. Med. activation).

Under these biomechanical constraints, the optimal control policy predicts physiological adaptations without exhaustive human EMG data. During the powerlift (Fig. 8), increasing the barbell weight to 10kg shifts the strategy, redistributing the load to deep trunk stabilizers and lower-body support, aligning with clinical force-transfer mechanics [15, 30]. Furthermore, this emergent fidelity is bounded by anatomical resolution (Fig. 9). Expanding to the 700-muscle model, the agent utilizes newly available axial stabilizers (deep trunk, neck, and scapular muscles) [117, 104] to optimize the lift [16, 68] (see detailed muscle activation heatmaps in App. E, Fig. 15 and Fig. 16). Ultimately, under high-fidelity anatomical constraints, reward-driven optimization converges on human-like motor control to counteract environmental forces.

IV. Residual RL Policy

Motivation. The preceding takeaways reveal complementary limitations of reward-only learning and reference-based imitation for full-body MSK control. Reward-based exploration in the raw 416-dimensional muscle-action space is difficult without embodiment-aware structure (**T1**), whereas imitation priors provide stable kinematic regularities but become brittle when contact timing, terrain height, or impact dynamics deviate from the reference distribution (**T2**). Moreover, task success alone is insufficient: physiology-oriented diagnostics such as EMG-envelope alignment, activation smoothness, and impact forces reveal differences in internal muscle-use strategies (**T3**). These observations motivate a residual adaptation strategy that preserves the stabilizing structure of an imitation prior while learning bounded, task-driven corrections for contact-rich and physiology-sensitive behaviors.

Method. We introduce an enhanced residual learning architecture (Fig. 10) that bridges the gap between imitation-based priors and reward-driven environmental adaptation. Specifically, a frozen imitation prior policy (π_{base}) [51] generates stabilizing base muscle excitations $\mathbf{a}_t^{\text{base}}$, while a trainable 3-layer MLP residual network (π_{res}) outputs targeted corrections $\mathbf{a}_t^{\text{res}}$. To preserve the fundamental synergistic patterns while enabling aggressive exploration, we disable standard kinematic penalties, retain only a minimal L2 regularization ($c_{\text{res}} = -0.02$) on the residual output, and bound the final excitation using a dynamic control ratio α: $\mathbf{a}_t^{\text{full}} = \text{clip}(\mathbf{a}_t^{\text{base}} + \alpha \cdot \text{clip}(\mathbf{a}_t^{\text{res}}, -1.0, 1.0), -1.0, 1.0)$. Crucially, we augment the

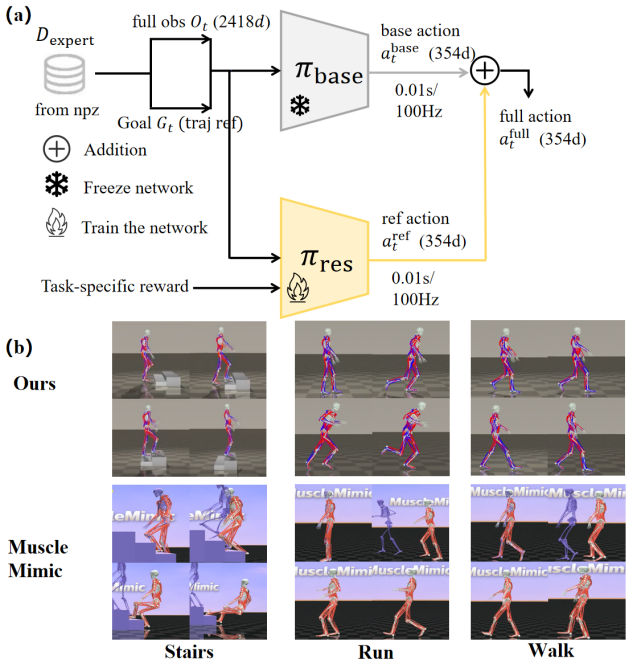

Fig. 10: **Residual RL Architecture and Performance.** (a) Overview of the residual framework, where the final action is the sum of a frozen imitation prior policy [51] and a task-optimized residual network. (b) Visual comparison of **stair-climbing**, **running** and **walking** gaits. Blue skeletons denote the reference imitation prior; colored trajectories illustrate the residual policy adaptations.

standard residual framework with task-aware tracking mechanisms and specialized reward structures (detailed in App. J):

- **Elastic Trajectory Pacing & Z-Agnostic Tracking (Stairs):** To prevent fatal error accumulation when step-

TABLE II: Comparison on the **walk** task.

Method	SR(%)	Pearson correlation r ↑		
		SoleusMed.	Bic.Fem.	Semitend.
MuscleMimic [51]	**100**	0.71	0.69	0.63
Ours	**100**	**0.75**	**0.80**	**0.67**

TABLE III: Comparison on the **stairs climbing** task.

Method	SR (%) ↑	Max Z-Height (m) ↑	Tracking Error (rad) ↓
MuscleMimic [51]	0.0	0.94	0.1968
Ours	**100**	**1.21**	**0.1079**

ping on physical obstacles, we decouple the reference phase clock from the simulation time. The trajectory index advances only if the planar root deviation is within a safe threshold ($d < 0.4$m); if the agent lags, the reference progression slows by 50%. Furthermore, we mask the Z-axis displacement in the imitation objective ($d_{\text{root}} = \|\mathbf{p}_{xy} - \hat{\mathbf{p}}_{xy}\|_2^2$) and introduce a progress reward based on velocity projected to future reference states (1.0s ahead), allowing the agent to actively step onto elevated platforms without suffering premature termination.

- **Phase-Dependent Authority & Metabolic Smoothing (Walk):** To eliminate high-frequency, non-functional muscle twitches, we reinstate a smoothed metabolic penalty: $p_{\text{met}} = -w_{\text{met}} \sum_j \left(a_{j,t}^2 + \lambda(a_{j,t} - a_{j,t-1})^2\right)$. Concurrently, we dynamically modulate the residual scaling factor α based on the gait phase ($\alpha_i(t) = \alpha_{\text{stance}} + (\alpha_{\text{swing}} - \alpha_{\text{stance}}) \cdot \mathcal{I}_{\text{swing}}(i,t)$). Assigning higher authority ($\alpha = 0.6$) to the swing leg grants flexibility for foot clearance, while lower authority ($\alpha = 0.2$) for the stance leg ensures synergistic weight-bearing stability.
- **Flight Propulsion & GRF Smoothing (Run):** High-speed running introduces aerial phases where reference timing often misaligns with simulated physics. We temporarily relax tracking penalties by 50% during the flight phase and apply a propulsive reward to encourage explosive toe-off. To manage high-speed impacts, we introduce a soft penalty for excessive vertical Ground Reaction Forces (GRF) exceeding 2.5× body weight, forcing the residual policy to learn active shock absorption strategies prior to touchdown.

Results. Tab. II demonstrates that our residual approach substantially improves physiological realism during the walking task, achieving closer alignment with ground-truth human EMG data. This improvement is directly attributed to our phase-dependent residual authority and metabolic smoothing, which force the network to trade off overly stiff kinematic tracking for natural, burst-like muscle recruitments.

TABLE IV: Comparison on the **high-impact run** task.

Method	SR (%)	Peak GRF (× Body Wt.) ↓	Tracking Error (rad) ↓
MuscleMimic [51]	**90**	6.40	0.2267
Ours	**100**	**6.07**	**0.1338**

Beyond flat-ground locomotion, our residual adaptation proves essential for complex environmental interactions. As detailed in Tab. III, pure imitation (MuscleMimic) fails completely to climb stairs (0% success rate) due to rigid height tracking penalties and timing desynchronization, stalling at a maximum Z-height of 0.94m. In contrast, our residual policy, guided by the elastic pacing and Z-agnostic objectives, successfully learns the necessary active lifting dynamics to navigate elevation changes (achieving a 100% success rate and 1.21m height) with significantly reduced tracking error (Fig. 10b). Furthermore, in the highly dynamic running task (Tab. IV), facilitated by GRF smoothing and flight-phase propulsion, our method effectively mitigates landing impacts. By actively reducing the peak ground reaction force (from 6.40× to 6.07× body weight), our method imitates the high-speed reference motion with greater stability and markedly reduced tracking error (0.1338 rad vs. 0.2267 rad) compared to the imitation-only baseline.

V. Limitations and Future Work

While MSK-Bench provides a comprehensive foundation for muscle-driven control, several limitations highlight avenues for future research (details in App. K). **First**, the suite focuses on gross motor skills, lacking fine-grained dexterous manipulation [12] due to the computational complexity of modeling hand biomechanics. **Second**, our policies are evaluated on isolated skills; real-world applications demand compositional, long-horizon behaviors, necessitating future explorations into hierarchical RL, skill discovery, and fatigue modeling. **Third**, due to the scarcity of comprehensive EMG datasets for dynamic movements, rigorous physiological validation is currently restricted to walking. **Finally**, bridging the sim-to-real gap remains a crucial next step. Deploying these policies requires overcoming unmodeled hardware dynamics, achievable either through tendon-driven robots or by mapping virtual muscle compliance onto torque-driven humanoids via whole-body impedance control.

VI. Conclusion

In this work, we introduced **MSK-Bench**, a full-body musculoskeletal benchmark for studying muscle-actuated humanoid control. Built on a highly overactuated 416-muscle humanoid, MSK-Bench provides 22 tasks spanning postural stabilization, locomotion, and contact-rich environmental interaction, together with a unified evaluation protocol covering task success, robustness, activation cost, joint smoothness, and EMG-envelope similarity. Our experiments reveal that full-body MSK control benefits from embodiment-aware exploration and action-structured representations, while imitation priors improve reference-compatible skills but become brittle under contact-rich terrain mismatch. We further demonstrate residual adaptation as a strong baseline that preserves the stabilizing benefits of imitation priors while correcting contact, timing, and muscle-use limitations. We hope MSK-Bench will provide a useful testbed for future research on physiologically grounded whole-body control.

References

[1] Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

[2] Firas Al-Hafez, Guoping Zhao, Jan Peters, and Davide Tateo. Locomujoco: A comprehensive imitation learning benchmark for locomotion. *arXiv preprint arXiv:2311.02496*, 2023.

[3] Minttu Alakuijala, Gabriel Dulac-Arnold, Julien Mairal, Jean Ponce, and Cordelia Schmid. Residual reinforcement learning from demonstrations. *arXiv preprint arXiv:2106.08050*, 2021.

[4] Lars Ankile, Zhenyu Jiang, Rocky Duan, Guanya Shi, Pieter Abbeel, and Anusha Nagabandi. Residual off-policy rl for finetuning behavior cloning policies. *arXiv preprint arXiv:2509.19301*, 2025.

[5] Sivakumar Balasubramanian, Alejandro Melendez-Calderon, Agnes Roby-Brami, and Etienne Burdet. On the analysis of movement smoothness. *Journal of neuroengineering and rehabilitation*, 12(1):112, 2015.

[6] Cameron Berg, Vittorio Caggiano, and Vikash Kumar. Sar: Generalization of physiological agility and dexterity via synergistic action representation. *Autonomous Robots*, 48(8):28, 2024.

[7] Junyo Boo, Dongwook Seo, Minseung Kim, and Seungbum Koo. Comprehensive human locomotion and electromyography dataset: Gait120. *Scientific Data*, 12 (1):1023, 2025.

[8] Simon Bouisset and Manh-Cuong Do. Posture, dynamic stability, and voluntary movement. *Neurophysiologie Clinique/Clinical Neurophysiology*, 38(6):345–362, 2008.

[9] Eadric Bressel, Jeffrey M Willardson, Brennan Thompson, and Fabio E Fontana. Effect of instruction, surface stability, and load intensity on trunk muscle activity. *Journal of Electromyography and Kinesiology*, 19(6): e500–e504, 2009.

[10] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[11] Vittorio Caggiano, Huawei Wang, Guillaume Durandau, Massimo Sartori, and Vikash Kumar. Myosuite–a contact-rich simulation suite for musculoskeletal motor control. *arXiv preprint arXiv:2205.13600*, 2022.

[12] Vittorio Caggiano, Sudeep Dasari, and Vikash Kumar. Myodex: a generalizable prior for dexterous manipulation. In *International Conference on Machine Learning*, pages 3327–3346. PMLR, 2023.

[13] Vittorio Caggiano, Guillaume Durandau, Seungmoon Song, Chun Kwang Tan, Huiyi Wang, Balint Hodossy, Pierre Schumacher, Letizia Gionfrida, Massimo Sartori, and Vikash Kumar. Myochallenge 2024: Physiological dexterity and agility in bionic humans. In *NeurIPS 2024 Competition Track*, 2024.

[14] Vittorio Caggiano, Cheryl Wang, Chun Kwang Tan, Balint K Hodossy, Shirui Lyu, Massimo Sartori, Seungmoon Song, Letizia Gionfrida, Guillaume Durandau, and Vikash Kumar. Myochallenge 2025: Towards human athletic intelligence. In *In NeuroIPS 2024 Competition Track*. 2025.

[15] Joaquin Calatayud, Juan C Colado, Fernando Martin, José Casaña, Markus D Jakobsen, and Lars L Andersen. Core muscle activity during the clean and jerk lift with barbell versus sandbags and water bags. *International journal of sports physical therapy*, 10(6):803, 2015.

[16] Giuseppe Coratella, Gianpaolo Tornatore, Stefano Longo, Fabio Esposito, and Emiliano Cè. Front vs back and barbell vs machine overhead press: an electromyographic analysis and implications for resistance training. *Frontiers in physiology*, 13:825880, 2022.

[17] R James Cotton. Kintwin: Imitation learning with torque and muscle driven biomechanical models enables precise replication of able-bodied and impaired movement from markerless motion capture. *arXiv preprint arXiv:2505.13436*, 2025.

[18] Roy D Crowninshield and Richard A Brand. A physiologically based criterion of muscle force prediction in locomotion. *Journal of biomechanics*, 14(11):793–801, 1981.

[19] Michael Damsgaard, John Rasmussen, Søren Tørholm Christensen, Egidijus Surma, and Mark De Zee. Analysis of musculoskeletal systems in the anybody modeling system. *Simulation Modelling Practice and Theory*, 14 (8):1100–1111, 2006.

[20] Andrea d'Avella, Philippe Saltiel, and Emilio Bizzi. Combinations of muscle synergies in the construction of a natural motor behavior. *Nature neuroscience*, 6(3): 300–308, 2003.

[21] DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026.

[22] Scott L Delp, Frank C Anderson, Allison S Arnold, Peter Loan, Ayman Habib, Chand T John, Eran Guendelman, and Darryl G Thelen. Opensim: open-source software to create and analyze dynamic simulations of movement. *IEEE transactions on biomedical engineering*, 54(11):1940–1950, 2007.

[23] Siddharth Desai, Haresh Karnan, Josiah P Hanna, Garrett Warnell, and Peter Stone. Stochastic grounded action transformation for robot learning in simulation. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6106–6111. IEEE, 2020.

[24] Michael H Dickinson, Claire T Farley, Robert J Full, MAR Koehl, Rodger Kram, and Steven Lehman. How animals move: an integrative view. *science*, 288(5463): 100–106, 2000.

[25] Kenji Doma, Glen B Deakin, and Kevin F Ness. Kinematic and electromyographic comparisons between

chin-ups and lat-pull down exercises. *Sports biomechanics*, 12(3):302–313, 2013.

[26] Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. In *International conference on machine learning*, pages 1329–1338. PMLR, 2016.

[27] Yusen Feng, Xiyan Xu, and Libin Liu. Musclevae: Model-based controllers of muscle-actuated characters. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023.

[28] Tamar Flash and Neville Hogan. The coordination of arm movements: an experimentally confirmed mathematical model. *Journal of neuroscience*, 5(7):1688–1703, 1985.

[29] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

[30] Stephan Geisler, Tim Havers, Eduard Isenmann, Jonas Schulze, Leonie K Lourens, Jannik Nowak, Steffen Held, and G Gregory Haff. Effects of expertise on muscle activity during the hang power clean and hang power snatch compared to snatch and clean pulls–an explorative analysis. *Journal of Sports Science & Medicine*, 22(4):778, 2023.

[31] Selim Gilon, Emily Y Miller, and Scott D Uhlrich. Opencap monocular: 3d human kinematics and musculoskeletal dynamics from a single smartphone video. *arXiv preprint arXiv:2603.24733*, 2026.

[32] Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, et al. Maniskill2: A unified benchmark for generalizable manipulation skills. *arXiv preprint arXiv:2302.04659*, 2023.

[33] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. Pmlr, 2018.

[34] Samuel R Hamner, Ajay Seth, and Scott L Delp. Muscle contributions to propulsion and support during running. *Journal of biomechanics*, 43(14):2709–2716, 2010.

[35] Lijun Han, Long Cheng, Muyuan Ma, Feiyang Yang, and Houcheng Li. Muscle synergy-guided reinforcement learning for embodied musculoskeletal motion skill learning. *IEEE Transactions on Biomedical Engineering*, 2026.

[36] Kaibo He, Chenhui Zuo, Chengtian Ma, and Yanan Sui. Dynsyn: Dynamical synergistic representation for efficient learning and control in overactuated embodied systems. *arXiv preprint arXiv:2407.11472*, 2024.

[37] Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. *arXiv preprint arXiv:2406.08858*, 2024.

[38] Jakob Hollenstein, Sayantan Auddy, Matteo Saveriano, Erwan Renaudo, and Justus Piater. Action noise in off-policy deep reinforcement learning: Impact on exploration and performance. *arXiv preprint arXiv:2206.03787*, 2022.

[39] Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J Davison. Rlbench: The robot learning benchmark and learning environment. ieee robotics and automation letters 5, 2 (2020), 3019–3026, 2020.

[40] Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Fei-Fei Li, Anima Anandkumar, Yuke Zhu, and Linxi Fan. Vima: General robot manipulation with multimodal prompts. *arXiv.org*, 2023. ISSN 2331-8422.

[41] Tobias Johannink, Shikhar Bahl, Ashvin Nair, Jianlan Luo, Avinash Kumar, Matthias Loskyll, Juan Aparicio Ojea, Eugen Solowjow, and Sergey Levine. Residual reinforcement learning for robot control. In *2019 international conference on robotics and automation (ICRA)*, pages 6023–6029. IEEE, 2019.

[42] Erika Jonsson, Åke Seiger, and Helga Hirschfeld. One-leg stance in healthy young and elderly adults: a measure of postural steadiness? *Clinical biomechanics*, 19 (7):688–694, 2004.

[43] Kento Kawaharazuka, Kei Okada, and Masayuki Inaba. Characteristics, management, and utilization of muscles in musculoskeletal humanoids: Empirical study on kengoro and musashi. *Advanced Intelligent Systems*, page e202500741, 2026.

[44] Łukasz Kidziński, Sharada P Mohanty, Carmichael F Ong, Jennifer L Hicks, Sean F Carroll, Sergey Levine, Marcel Salathé, and Scott L Delp. Learning to run challenge: Synthesizing physiologically accurate motion using deep reinforcement learning. In *The NIPS'17 Competition: Building Intelligent Systems*, pages 101–120. Springer, 2018.

[45] Łukasz Kidziński, Sharada Prasanna Mohanty, Carmichael F Ong, Zhewei Huang, Shuchang Zhou, Anton Pechenko, Adam Stelmaszczyk, Piotr Jarosik, Mikhail Pavlov, Sergey Kolesnikov, et al. Learning to run challenge solutions: Adapting reinforcement learning methods for neuromusculoskeletal environments. In *The NIPS'17 Competition: Building Intelligent Systems*, pages 121–153. Springer, 2018.

[46] Łukasz Kidziński, Carmichael Ong, Sharada Prasanna Mohanty, Jennifer Hicks, Sean Carroll, Bo Zhou, Hongsheng Zeng, Fan Wang, Rongzhong Lian, Hao Tian, et al. Artificial intelligence for prosthetics: Challenge solutions. In *The NeurIPS'18 Competition: From Machine Learning to Intelligent Conversations*, pages 69–128. Springer, 2019.

[47] Yuxuan Kuang, Haoran Geng, Amine Elhafsi, Tan-Dzung Do, Pieter Abbeel, Jitendra Malik, Marco Pavone, and Yue Wang. Skillblender: Towards versa-

tile humanoid whole-body loco-manipulation via skill blending. *arXiv preprint arXiv:2506.09366*, 2025.

[48] Vikash Kumar, Rutav Shah, Gaoyue Zhou, Vincent Moens, Vittorio Caggiano, Abhishek Gupta, and Aravind Rajeswaran. Robohive: A unified framework for robot learning. *Advances in Neural Information Processing Systems*, 36:44323–44340, 2023.

[49] Arthur D Kuo and J Maxwell Donelan. Dynamic principles of gait and their clinical implications. *Physical therapy*, 90(2):157–174, 2010.

[50] Mark L Latash. Muscle coactivation: definitions, mechanisms, and functions. *Journal of neurophysiology*, 120 (1):88–104, 2018.

[51] Chengkun Li, Cheryl Wang, Bianca Ziliotto, Merkourios Simos, Jozsef Kovecses, Guillaume Durandau, and Alexander Mathis. Towards embodied ai with musclemimic: Unlocking full-body musculoskeletal motor learning at scale. *arXiv preprint arXiv:2603.25544*, 2026.

[52] Toru Lin, Shuying Deng, Zhao-Heng Yin, Pieter Abbeel, and Jitendra Malik. How to peel with a knife: Aligning fine-grained manipulation with human preference. *arXiv preprint arXiv:2603.03280*, 2026.

[53] Xingyu Lin, Yufei Wang, Jake Olkin, and David Held. Softgym: Benchmarking deep reinforcement learning for deformable object manipulation. In *Conference on Robot Learning*, pages 432–448. PMLR, 2021.

[54] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36: 44776–44791, 2023.

[55] May Q Liu, Frank C Anderson, Michael H Schwartz, and Scott L Delp. Muscle contributions to support and progression over a range of walking speeds. *Journal of biomechanics*, 41(15):3243–3252, 2008.

[56] Yun Liu, Bowen Yang, Licheng Zhong, He Wang, and Li Yi. Mimicking-bench: A benchmark for generalizable humanoid-scene interaction learning via human mimicking. *arXiv preprint arXiv:2412.17730*, 2024.

[57] Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Jim Fan, et al. Eureka: Human-level reward design via coding large language models. In *International conference on learning Representations*, volume 2024, pages 26516–26560, 2024.

[58] Jean Massion. Movement, posture and equilibrium: interaction and coordination. *Progress in neurobiology*, 38(1):35–56, 1992.

[59] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334, 2022.

[60] Florian Michaud, Mario Lamas, Urbano Lugrís, and Javier Cuadrado. A fair and emg-validated comparison of recruitment criteria, musculotendon models and muscle coordination strategies, for the inverse-dynamics based optimization of muscle forces during gait. *Journal of neuroengineering and rehabilitation*, 18(1):17, 2021.

[61] H Moffet, CL Richards, F Malouin, and G Bravo. Load-carrying during stair ascent: a demanding functional test. *Gait & Posture*, 1(1):35–44, 1993.

[62] Yao Mu, Tianxing Chen, Zanxin Chen, Shijia Peng, Zhiqian Lan, Zeyu Gao, Zhixuan Liang, Qiaojun Yu, Yude Zou, Mingkun Xu, et al. Robotwin: Dual-arm robot benchmark with generative digital twins. In *Proceedings of the computer vision and pattern recognition conference*, pages 27649–27660, 2025.

[63] Yuxuan Mu, Ziyu Zhang, Yi Shi, Minami Matsumoto, Kotaro Imamura, Guy Tevet, Chuan Guo, Michael Taylor, Chang Shu, Pengcheng Xi, et al. Smp: Reusable score-matching motion priors for physics-based character control. *arXiv preprint arXiv:2512.03028*, 2025.

[64] Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. *arXiv preprint arXiv:2406.02523*, 2024.

[65] Soroush Nasiriany, Sepehr Nasiriany, Abhiram Maddukuri, and Yuke Zhu. Robocasa365: A large-scale simulation framework for training and benchmarking generalist robots. *arXiv preprint arXiv:2603.04356*, 2026.

[66] Richard R Neptune and Arian Vistamehr. Dynamic balance during human movement: measurement and control mechanisms. *Journal of Biomechanical Engineering*, 141(7):070801, 2019.

[67] Karl M Newell. What are fundamental motor skills and what is fundamental about them? *Journal of Motor Learning and Development*, 8(2):280–314, 2020.

[68] Ashish D Nimbarte. Risk of neck musculoskeletal disorders among males and females in lifting exertions. *International Journal of Industrial Ergonomics*, 44(2): 253–259, 2014.

[69] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions On Graphics (TOG)*, 37(4): 1–14, 2018.

[70] Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 3803–3810. IEEE, 2018.

[71] Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (ToG)*, 40(4):1–20, 2021.

[72] Xue Bin Peng, Yunrong Guo, Lina Halper, Sergey Levine, and Sanja Fidler. Ase: Large-scale reusable adversarial skill embeddings for physically simulated

characters. *ACM Transactions On Graphics (TOG)*, 41 (4):1–17, 2022.

[73] Ilija Radosavovic, Tete Xiao, Bike Zhang, Trevor Darrell, Jitendra Malik, and Koushil Sreenath. Real-world humanoid locomotion with reinforcement learning. *Science Robotics*, 9(89):eadi9579, 2024.

[74] Krishan Rana, Ming Xu, Brendan Tidd, Michael Milford, and Niko Sünderhauf. Residual skill policies: Learning an adaptable skill-based action space for reinforcement learning for robotics. In *Conference on Robot Learning*, pages 2095–2104. PMLR, 2023.

[75] Davis Rempe, Mathis Petrovich, Ye Yuan, Haotian Zhang, Xue Bin Peng, Yifeng Jiang, Tingwu Wang, Umar Iqbal, David Minor, Michael de Ruyter, et al. Kimodo: Scaling controllable human motion generation. *arXiv preprint arXiv:2603.15546*, 2026.

[76] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[77] Pierre Schumacher, Daniel Häufle, Dieter Büchler, Syn Schmitt, and Georg Martius. Dep-rl: Embodied exploration for reinforcement learning in overactuated and musculoskeletal systems. *arXiv preprint arXiv:2206.00484*, 2022.

[78] Jessica C Selinger, Shawn M O'Connor, Jeremy D Wong, and J Maxwell Donelan. Humans can continuously optimize energetic cost during walking. *Current Biology*, 25(18):2452–2456, 2015.

[79] Ajay Seth, Michael Sherman, Jeffrey A Reinbolt, and Scott L Delp. Opensim: a musculoskeletal modeling and simulation framework for in silico investigations and exchange. *Procedia Iutam*, 2:212–232, 2011.

[80] Ajay Seth, Jennifer L Hicks, Thomas K Uchida, Ayman Habib, Christopher L Dembia, James J Dunne, Carmichael F Ong, Matthew S DeMers, Apoorva Rajagopal, Matthew Millard, et al. Opensim: Simulating musculoskeletal dynamics and neuromuscular control to study human and animal movement. *PLoS computational biology*, 14(7):e1006223, 2018.

[81] Carmelo Sferrazza, Dun-Ming Huang, Xingyu Lin, Youngwoon Lee, and Pieter Abbeel. Humanoidbench: Simulated humanoid benchmark for whole-body locomotion and manipulation. *arXiv preprint arXiv:2403.10506*, 2024.

[82] Tom Silver, Kelsey Allen, Josh Tenenbaum, and Leslie Kaelbling. Residual policy learning. *arXiv preprint arXiv:1812.06298*, 2018.

[83] Merkourios Simos, Alberto Silvio Chiappa, and Alexander Mathis. Reinforcement learning-based motion imitation for physiologically plausible musculoskeletal motor control. *arXiv e-prints*, pages arXiv–2503, 2025.

[84] Ronald L Snarr, Ashleigh V Hallmark, Jason C Casey, and Michael R Esco. Electromyographical comparison of a traditional, suspension device, and towel pull-up. *Journal of Human Kinetics*, 58:5, 2017.

[85] Seungmoon Song, Łukasz Kidziński, Xue Bin Peng, Carmichael Ong, Jennifer Hicks, Sergey Levine, Christopher G Atkeson, and Scott L Delp. Deep reinforcement learning for modeling human locomotion control in neuromechanical simulation. *Journal of neuroengineering and rehabilitation*, 18(1):126, 2021.

[86] Ian AF Stokes, Mack G Gardner-Morse, and Sharon M Henry. Abdominal muscle activation increases lumbar spinal stability: analysis of contributions of different muscle groups. *Clinical biomechanics*, 26(8):797–803, 2011.

[87] Zhenguo Sun, Yibo Peng, Yuan Meng, Xukun Li, Bo-Sheng Huang, Zhenshan Bing, Xinlong Wang, and Alois Knoll. Robotdancing: Residual-action reinforcement learning enables robust long-horizon humanoid motion tracking. *arXiv preprint arXiv:2509.20717*, 2025.

[88] Jie Tan, Tingnan Zhang, Erwin Coumans, Atil Iscen, Yunfei Bai, Danijar Hafner, Steven Bohez, and Vincent Vanhoucke. Sim-to-real: Learning agile locomotion for quadruped robots. *arXiv preprint arXiv:1804.10332*, 2018.

[89] Stone Tao, Fanbo Xiang, Arth Shukla, Yuzhe Qin, Xander Hinrichsen, Xiaodi Yuan, Chen Bao, Xinsong Lin, Yulin Liu, Tse-kai Chan, et al. Maniskill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai. *arXiv preprint arXiv:2410.00425*, 2024.

[90] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

[91] Chen Tessler, Yoni Kasten, Yunrong Guo, Shie Mannor, Gal Chechik, and Xue Bin Peng. Calm: Conditional adversarial latent models for directable virtual characters. In *ACM SIGGRAPH 2023 conference proceedings*, pages 1–9, 2023.

[92] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017.

[93] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE, 2012.

[94] Scott D Uhlrich, Antoine Falisse, Łukasz Kidziński, Julie Muccini, Michael Ko, Akshay S Chaudhari, Jennifer L Hicks, and Scott L Delp. Opencap: Human movement dynamics from smartphone videos. *PLoS computational biology*, 19(10):e1011462, 2023.

[95] Francisco J Valero-Cuevas, BA Cohn, HF Yngvason, and Emily L Lawrence. Exploring the high-dimensional structure of muscle redundancy via subject-specific and

generic musculoskeletal models. *Journal of biomechanics*, 48(11):2887–2896, 2015.

[96] Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pages 1723–1736. PMLR, 2023.

[97] Cheryl Wang, Chun Kwang Tan, Balint K. Hodossy, Eric Lyu, Jun Guo, Wentao Zhao, Huaping Liu, Chengkun Li, Merkourios Simos, Bianca Ziliotto, Alexander Mathis, Siyuan Liu, Jiahao Chen, Shanlin Zhong, Bo Jiang, Ci Song, Yaoye Zhu, Chenhui Zuo, Yanan Sui, Mohamed Irfan Refai, Massimo Sartori, Guillaume Durandau, Vikash Kumar, and Vittorio Caggiano. Myochallenge 2025: A new benchmark for human athletic intelligence, 2026. URL https://arxiv.org/abs/2605.15650.

[98] Huawei Wang, Vittorio Caggiano, Guillaume Durandau, Massimo Sartori, and Vikash Kumar. Myosim: Fast and physiologically realistic mujoco models for musculoskeletal and exoskeletal studies. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 8104–8111. IEEE, 2022.

[99] Huiyi Wang, Chun Kwang Tan, Balint Hodossy, Shirui Lyu, Pierre Schumacher, James Heald, Kai Biegun, Samo Hromadka, Maneesh Sahani, Gunwoo Park, et al. Myochallenge 2024: A new benchmark for physiological dexterity and agility in bionic humans. *Advances in Neural Information Processing Systems*, 38, 2026.

[100] Lirui Wang, Yiyang Ling, Zhecheng Yuan, Mohit Shridhar, Chen Bao, Yuzhe Qin, Bailin Wang, Huazhe Xu, and Xiaolong Wang. Gensim: Generating robotic simulation tasks via large language models. In *International Conference on Learning Representations*, volume 2024, pages 4890–4924, 2024.

[101] Tingwu Wang, Olivier Dionne, Michael De Ruyter, David Minor, Davis Rempe, Kaifeng Zhao, Mathis Petrovich, Ye Yuan, Chenran Li, Zhengyi Luo, et al. Motionbricks: Scalable real-time motions with modular latent generative model and smart primitives. *arXiv preprint arXiv:2604.24833*, 2026.

[102] Yunyue Wei, Shanning Zhuang, Vincent Zhuang, and Yanan Sui. Motion control of high-dimensional musculoskeletal systems with hierarchical model-based planning. *arXiv preprint arXiv:2505.08238*, 2025.

[103] Yunyue Wei, Chenhui Zuo, and Yanan Sui. Scalable exploration for high-dimensional continuous control via value-guided flow. *arXiv preprint arXiv:2601.19707*, 2026.

[104] Yunyue Wei, Chenhui Zuo, Shanning Zhuang, Haixin Gong, Yaming Liu, and Yanan Sui. Scaling whole-body human musculoskeletal behavior emulation for specificity and diversity. *arXiv preprint arXiv:2603.29332*, 2026.

[105] Yuxin Wen, Qing Shuai, Di Kang, Jing Li, Cheng Wen, Yue Qian, Ningxin Jiao, Changhai Chen, Weijie Chen, Yiran Wang, et al. Hy-motion 1.0: Scaling flow matching models for text-to-motion generation. *arXiv preprint arXiv:2512.23464*, 2025.

[106] Tianbao Xie, Siheng Zhao, Chen Wu, Yitao Liu, Qian Luo, Victor Zhong, Yanchao Yang, and Tao Yu. Text2reward: Reward shaping with language models for reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pages 35663–35699, 2024.

[107] Pei Xu, Yufei Ye, Shuchun Sun, Yu Ding, Elizabeth Schumann, and C Karen Liu. Music: Learning muscle-driven dexterous hand control. *arXiv preprint arXiv:2604.23886*, 2026.

[108] Xiaomeng Xu, Yifan Hou, Zeyi Liu, and Shuran Song. Compliant residual dagger: Improving real-world contact-rich manipulation with human corrections. *Advances in Neural Information Processing Systems*, 38: 139559–139581, 2026.

[109] Yashuai Yan, Esteve Valls Mascaro, Tobias Egle, and Dongheui Lee. I-ctrl: Imitation to control humanoid robots through constrained reinforcement learning. *arXiv preprint arXiv:2405.08726*, 2024.

[110] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pages 1094–1100. PMLR, 2020.

[111] Wenhao Yu, Nimrod Gileadi, Chuyuan Fu, Sean Kirmani, Kuang-Huei Lee, Montse Gonzalez Arenas, Hao-Tien Lewis Chiang, Tom Erez, Leonard Hasenclever, Jan Humplik, et al. Language to rewards for robotic skill synthesis. *arXiv preprint arXiv:2306.08647*, 2023.

[112] Felix E Zajac, Richard R Neptune, and Steven A Kautz. Biomechanics and muscle coordination of human walking: Part i: Introduction to concepts, power transfer, dynamics and simulations. *Gait & posture*, 16(3):215–232, 2002.

[113] Felix E Zajac, Richard R Neptune, and Steven A Kautz. Biomechanics and muscle coordination of human walking: part ii: lessons from dynamical simulations and clinical implications. *Gait & posture*, 17(1):1–17, 2003.

[114] Haozhuo Zhang, Jingkai Sun, Michele Caprio, Jian Tang, Shanghang Zhang, Qiang Zhang, and Wei Pan. Lhm-humanoid: Learning a unified policy for long-horizon humanoid whole-body loco-manipulation in diverse messy environments, 2026. URL https://arxiv.org/abs/2508.16943.

[115] Siheng Zhao, Yanjie Ze, Yue Wang, C Karen Liu, Pieter Abbeel, Guanya Shi, and Rocky Duan. Resmimic: From general motion tracking to humanoid whole-body loco-manipulation via residual learning. *arXiv preprint arXiv:2510.05070*, 2025.

[116] Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Kevin Lin, Abhiram Maddukuri, Soroush Nasiriany, and Yifeng Zhu. robo-

suite: A modular simulation framework and benchmark for robot learning. *arXiv preprint arXiv:2009.12293*, 2020.

[117] Chenhui Zuo, Kaibo He, Jing Shao, and Yanan Sui. Self model for embodied intelligence: Modeling full-body human musculoskeletal system and locomotion control with hierarchical low-dimensional representation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13062–13069. IEEE, 2024.

APPENDIX

This appendix provides supplementary technical details, mathematical derivations, implementation descriptions, and extended experimental results to support the main findings of the paper. The content is organized as follows:

- **App. A: Task Specification.** Defines the common kinematic variables, orientation quantities, control symbols, reward terms, task settings, observations, initialization protocols, and termination criteria for the full musculoskeletal humanoid benchmark.
- **App. B: Evaluation Details.** Describes the evaluation metrics used in the benchmark, including success rate, robustness under action noise, observation noise and dynamics randomization, muscle activation energy, joint smoothness, peak efficiency steps, and EMG similarity.
- **App. C: Detailed Descriptions of Reward-Based RL Baselines.** Provides detailed explanations of the pure reinforcement learning baselines, including PPO, SAC, Dep-RL, and DynSyn-SAC, with emphasis on their exploration mechanisms and suitability for high-dimensional musculoskeletal control.
- **App. D: Detailed Reward Curves for Reward-Based RL Baselines.** Presents the complete reward curves for all 22 benchmark tasks, enabling a detailed comparison of convergence behavior and training stability across different RL baselines.
- **App. E: Detailed Muscle Activation Analysis.** Provides time-step-resolved muscle activation heatmaps and biomechanical analysis for the 416-muscle and 700-muscle models, highlighting how anatomical fidelity affects load distribution and recruitment strategies.
- **App. F: Agentic Reward Tuning.** Describes the proposed closed-loop reward refinement mechanism, where an external reasoning module periodically adjusts reward weights according to recent gait quality, posture stability, and task-progress statistics.
- **App. G: Latent-Action RL.** Explains the expert-informed hard-mode action filtering framework, where a pretrained anatomical encoder-decoder constrains RL actions within a bounded neighborhood of biologically plausible muscle activations.
- **App. H: EMG Data Processing and Comparative Analysis.** Details the gait-cycle normalization, multi-cycle averaging, phase alignment, and Pearson-correlation-based comparison between simulated muscle activations and human EMG envelopes.
- **App. I: Muscle Recruitment Categorization.** Summarizes the anatomical keyword-based grouping of 416 muscle actuators into interpretable muscle families, allowing whole-body activation patterns to be analyzed at the functional group level.
- **App. J: Detailed Descriptions of Residual RL Policy.** Describes the residual reinforcement learning framework used to adapt the base tracking policy to challenging locomotion settings. This section details the residual action formulation, residual scaling strategy, Z-axis-agnostic imitation objective for stair navigation, elastic trajectory pacing, phase-dependent residual authority, metabolic-aware muscle regularization, and high-impact locomotion adaptations for running.
- **App. K: Limitations and Future Directions.** Discusses the current limitations of the benchmark and outlines future research directions, including dexterous manipulation, compositional and long-horizon task learning, broader physiological validation with EMG data, and sim-to-real transfer for physical musculoskeletal humanoid systems.

A. Task Specification

Before enumerating the environment details below, let us define common auxiliary variables, symbols, and regularizations that are broadly employed across the reward functions of multiple environments. The complete task suite is visually summarized in Figs. 11–13, where Fig. 11 presents whole-body stabilization tasks, Fig. 12 illustrates representative locomotion behaviors and Fig. 13 shows terrain and object interaction tasks.

Kinematics and Dynamics:

- $q, \dot{q}$ denote the proprioceptive joint positions and velocities of the robot. Specifically, q_{joints} denotes the subset of non-root joint angles, and $\bar{q}_{\text{hip}}, \bar{q}_{\text{knee}}$ represent the average angles of the left and right hip/knee joints.
- $\mathbf{p}_{(\cdot)}$ represents the 3D position of a specific body part or object in the global frame (e.g., $\mathbf{p}_{\text{pelvis}}, \mathbf{p}_{\text{foot,L}}$). The superscript xy (e.g., $\mathbf{p}^{xy}_{\text{pelvis}}$) denotes its 2D horizontal projection.
- $\mathbf{v}_{(\cdot)}$ represents the 3D linear velocity vector. Specifically, $\mathbf{v}_{\text{CoM}}$ is the center-of-mass velocity, with $v_{x,\text{CoM}}, v_{y,\text{CoM}}, v_{z,\text{CoM}}$ being its scalar components.
- $z_{(\cdot)}$ denotes the vertical height of a specific body part or object from the ground (e.g., $z_{\text{CoM}}, z_{\text{pelvis}}, z_{\text{torso}}, z_{\text{head}}, z_{\text{foot}}$).
- $\Delta\mathbf{p}_{\text{feet}} = \mathbf{p}_{\text{foot,L}} - \mathbf{p}_{\text{foot,R}}$ denotes the relative 3D vector between the left and right feet. We use $d_{xy,\text{feet}}$ to denote the horizontal Euclidean separation between the feet, and $d_{x,\text{feet}}, d_{y,\text{feet}}, d_{z,\text{feet}}$ for its absolute scalar components.
- $\Delta z_{\text{leg}} = z_{\text{pelvis}} - \min(z_{\text{foot,L}}, z_{\text{foot,R}})$ defines the robot's vertical leg extension, primarily used to enforce an upright posture and prevent deep-squatting or kneeling exploits.

Orientations and Vectors:

- $\mathbf{u}_{\text{torso}}$ is the normalized 3D unit vector pointing from the robot's pelvis to its head, used to evaluate general core verticality.
- $\mathbf{u}_{\text{chest}}$ is the normalized unit vector pointing forward from the robot's chest, used to penalize hunching or twisting.
- $\mathbf{u}_{\text{fwd}}$ is the normalized forward-facing vector derived from the pelvis orientation.
- $\hat{\mathbf{x}} = [1,0,0]^\top, \hat{\mathbf{y}} = [0,1,0]^\top, \hat{\mathbf{z}} = [0,0,1]^\top$ are the global forward, lateral, and vertical unit vectors.

- $y_{\text{yaw,dev}}$ denotes the forward yaw angular deviation derived from the robot's pelvis orientation, utilized to restrict unintended turning.

Control, Environment, and Regularization:

- $a \in \mathbb{R}^M$ denotes the muscle activations at the current control step, with a_t and a_{t-1} specifying temporal steps.
- ϕ is the time-varying gait phase clock, represented in observations as $(\sin\phi, \cos\phi)$, with ω being its angular frequency.
- $\mathbf{z}_{\text{terrain}}$ is the local terrain observation grid representing relative ground heights.
- $\Delta\mathbf{h} = [\Delta x, h_{\text{target}}, 0]^\top$ denotes the relative 3D vector to an upcoming obstacle/hurdle.
- $\mathbb{I}_{\text{alive}}$ represents a constant survival bonus awarded at each step.
- $R_{\text{efficiency}} = \text{mean}(a^2)$ is the standard action regularization term applied across tasks to penalize excessive muscle force and promote energy-efficient behaviors. $\sum$penalties accumulates minor supplementary regularizations (e.g., joint limits).

The detailed reward functions, task settings, specific observations, and termination criteria across all 22 tasks in the Musculoskeletal Humanoid Benchmark are systematically summarized in Tables V to VII (Reward Functions) and Tables VIII to X (Setup and Termination).

B. Evaluation Details

In this section, we present a detailed quantitative evaluation of the benchmarked tasks. We systematically assess the performance, sample efficiency, and zero-shot robustness of various reinforcement learning baselines under increasingly difficult conditions. The detailed results for the Stabilization (Table XI), Locomotion (Table XII), and Interaction (Table XIII) series are summarized in the tables below.

1) Success Rate: The percentage of test episodes in which the agent fulfills the task-specific success condition without falling or violating predetermined physical soft-constraints is defined as the success rate [11, 69, 81]. Higher success rate indicates better task performance. Formally, for N evaluation episodes, the success rate is computed as:

$$\text{Success Rate} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\Big(\mathcal{C}_{\text{task}}^{(i)} \wedge \mathcal{C}_{\text{alive}}^{(i)} \wedge \mathcal{C}_{\text{safe}}^{(i)}\Big), \tag{1}$$

where $\mathbb{I}(\cdot)$ is the indicator function, $\mathcal{C}_{\text{task}}^{(i)}$ indicates whether the i-th episode meets the task-specific completion criterion, $\mathcal{C}_{\text{alive}}^{(i)}$ indicates whether the agent avoids falling, and $\mathcal{C}_{\text{safe}}^{(i)}$ indicates whether the agent stays within the predetermined physical soft-constraints.

For instance, in the walking task, an episode is deemed successful only if the agent survives for at least $T_{\min}$ and travels farther than a minimal goal distance d_{target}:

$$\text{Success Rate}_{\text{walk}} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\Bigg(\Big\|\mathbf{p}_{\text{end}}^{(i)} - \mathbf{p}_{\text{start}}^{(i)}\Big\|_2 \geq d_{\text{target}} \wedge T_{\text{steps}}^{(i)} \geq T_{\min}\Bigg), \tag{2}$$

where N is the total number of evaluation episodes, $\mathbb{I}(\cdot)$ is the indicator function, $\mathbf{p}_{\text{start}}^{(i)}$ and $\mathbf{p}_{\text{end}}^{(i)}$ represent the agent's starting and final planar pelvis or center-of-mass positions in the i-th episode, and $T_{\text{steps}}^{(i)}$ denotes the number of survived simulation steps.

2) Robustness: Action noise, observation noise, and dynamics randomization are introduced to assess robustness [36, 73] of the methods. It evaluates whether a policy can maintain balance, preserve task performance, or quickly recover its intended trajectory when facing stochastic perturbations and randomized dynamics [36, 73]. To comprehensively test robustness under Sim2Real-style perturbations [92, 70, 88, 1], we consider the following:

First, **action noise** mimics stochastic execution errors in motor-neural excitation signals or low-level actuator commands. Sim-to-real robot learning studies that model stochastic action effects and transition mismatch between simulation and the real world [23], together with continuous-control reinforcement learning studies that examine additive action noise [38], motivate such action-space perturbations. In our implementation, Gaussian noise is applied to the normalized action vector generated by the policy. Since the policy action space is bounded within $[-1, 1]$, the perturbed action is clipped to this valid action range:

$$\begin{aligned} \mathbf{u}_t' &= \text{clip}_{[-1,1]}\left(\mathbf{u}_t + \boldsymbol{\epsilon}_t^{\text{act}}\right), \\ \boldsymbol{\epsilon}_t^{\text{act}} &\sim \mathcal{N}\left(\mathbf{0}, \sigma_{\text{act}}^2\mathbf{I}\right), \end{aligned} \tag{3}$$

where $\mathbf{u}_t$ denotes the policy's original normalized action output, $\boldsymbol{\epsilon}_t^{\text{act}}$ denotes Gaussian action noise, σ_{act} controls the action noise magnitude, and $\mathbf{u}_t'$ denotes the noisy action applied to the environment.

Second, **observation noise** mimics sensing errors from external sensors such as IMUs or proprioceptive feedback such as muscle spindles and Golgi tendon organs. This perturbation is also consistent with prior sim-to-real studies that perturb or randomize sensory inputs to improve robustness under observation mismatch [92, 70, 1]. In our implementation, Gaussian noise is added to the normalized policy observation vector before it is passed to the policy network:

$$\begin{aligned} \tilde{\mathbf{s}}_t' &= \text{clip}_{[-20,20]}\left(\tilde{\mathbf{s}}_t + \boldsymbol{\epsilon}_t^{\text{obs}}\right), \\ \boldsymbol{\epsilon}_t^{\text{obs}} &\sim \mathcal{N}\left(\mathbf{0}, \sigma_{\text{obs}}^2\mathbf{I}\right), \end{aligned} \tag{4}$$

where $\tilde{\mathbf{s}}_t$ denotes the normalized observation vector before perturbation, $\boldsymbol{\epsilon}_t^{\text{obs}}$ denotes Gaussian observation noise, σ_{obs} controls the observation noise level, and $\tilde{\mathbf{s}}_t'$ denotes the perturbed normalized observation received by the policy. The clipping operation is applied element-wise to keep the noisy normalized observation within the predefined numerical range.

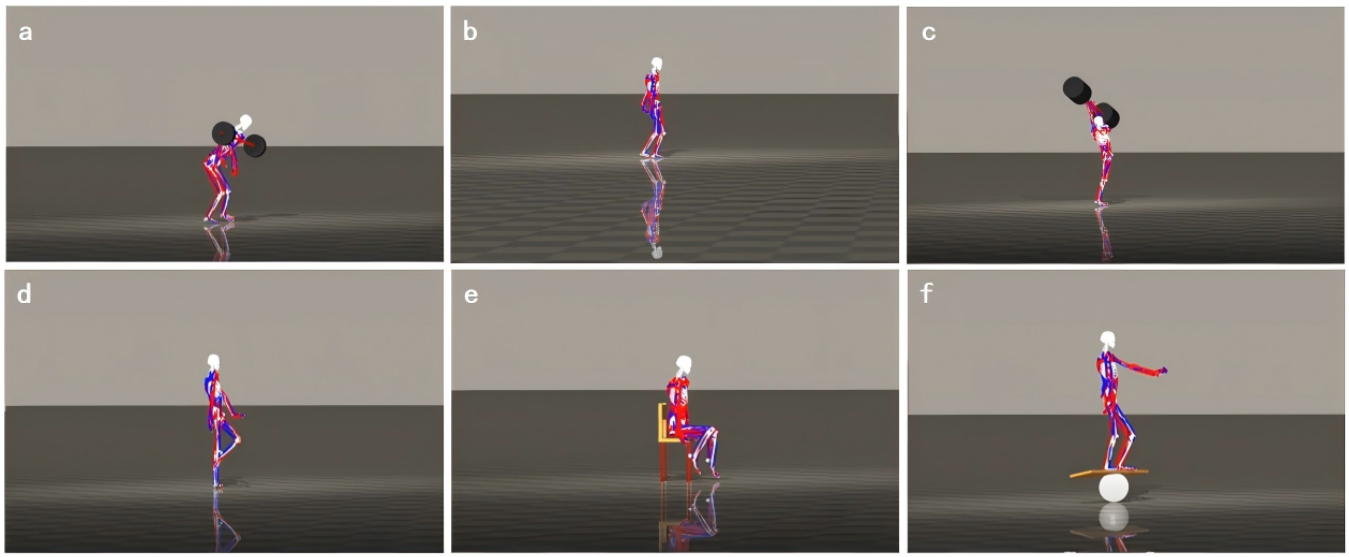

Fig. 11: **Illustration of six representative whole-body stabilization tasks.** (a) **squat**: The control policy must accurately track the dynamic vertical trajectories of the hip and knee based on a phase clock ϕ, while maintaining a strict center-of-mass balance lock between the two feet. A specially designed "landmine penalty" is introduced to immediately eliminate abnormal strategies that attempt to exploit the height-tracking reward by kneeling on both knees. (b) **stand**: The robot is required to maintain an absolutely static upright posture from an initial neutral pose. The key challenge is that the system must use high muscle stiffness to lock the joints, while even slight center-of-mass drift or foot sliding is severely penalized. The reward function is strongly biased toward absolute stillness and posture maintenance. (c) **powerlift**: To sustain the extreme load, the reward function explicitly encourages a biomechanically plausible forward-leaning hip-hinge posture through the R_{hinge} reward, thereby effectively recruiting the powerful posterior-chain muscle groups. Once dangerous lumbar collapse is detected, the episode is immediately terminated to ensure structural safety. (d) **Singlelegstand**: The robot must lift one leg into a "flamingo" posture and strictly align the center-of-mass projection with the support foot. The task severely penalizes lateral trunk lean, and after an initial physical tolerance period of 10 steps, the episode is immediately terminated once the lifted foot touches the ground. (e) **sit**: The robot is required to approach a chair and execute a precise sitting transition. To prevent cheating behaviors such as stiff-legged falling, the task enforces active knee flexion, requires the feet to be retracted beneath the support area, and introduces a "soft-landing" reward to limit pelvic descent velocity, ensuring a smooth motion. (f) **balance**: The robot is required to maintain dynamic balance on an unconstrained wobble board placed on a sphere. To overcome the common failure mode of falling backward, the task explicitly enforces a forward-leaning "fighting defensive stance", with a target pitch angle of $-0.15\,\text{rad}$, and requires precise center-of-mass control to keep the balance board level.

Third, **dynamics randomization** mimics individual variations in muscle physiology and uncertainty in muscle force capacity. To improve robustness against real-world dynamics mismatch [70, 88, 1], this setting follows the broader sim-to-real practice of randomizing physical parameters, such as mass, friction, actuator properties, and other dynamics-related quantities. At the beginning of each evaluation episode, we independently sample a random perturbation for each muscle and keep it fixed throughout the episode. Specifically, for the m-th muscle, we sample:

$$\begin{aligned} \delta_m &\sim \mathcal{U}\left(-\sigma_{\text{dyn}}, \sigma_{\text{dyn}}\right), \\ r_m &= 1 + \delta_m, \\ \text{Gain}'_m &= \text{Gain}_m \times r_m, \end{aligned} \tag{5}$$

where δ_m denotes the independently sampled perturbation for the m-th muscle, r_m denotes the corresponding multiplicative gain factor, Gain_m denotes the original maximum force-generation parameter of the m-th muscle, and Gain'_m denotes the randomized muscle gain used during evaluation. The parameter $0 \leq \sigma_{\text{dyn}} < 1$ controls the perturbation strength and ensures positive muscle gains. For example, $\sigma_{\text{dyn}} = 0.20$ means that each muscle's maximum force capacity is independently scaled by a random factor in $[0.8, 1.2]$.

Across these perturbation settings, we report the average reward, survival steps, and true success rate to quantify how reliably each policy maintains task performance under noise and randomized dynamics. Higher values of these metrics under perturbation indicate stronger robustness. When compared with the clean evaluation setting, smaller performance degradation indicates better robustness.

3) Muscle Activation Energy: Muscle activation energy [117, 104, 36, 78, 60] is used to quantify the average muscle activation cost by computing the average squared muscle activation of all underlying muscles across the whole body within a time window. This metric is used as a proxy

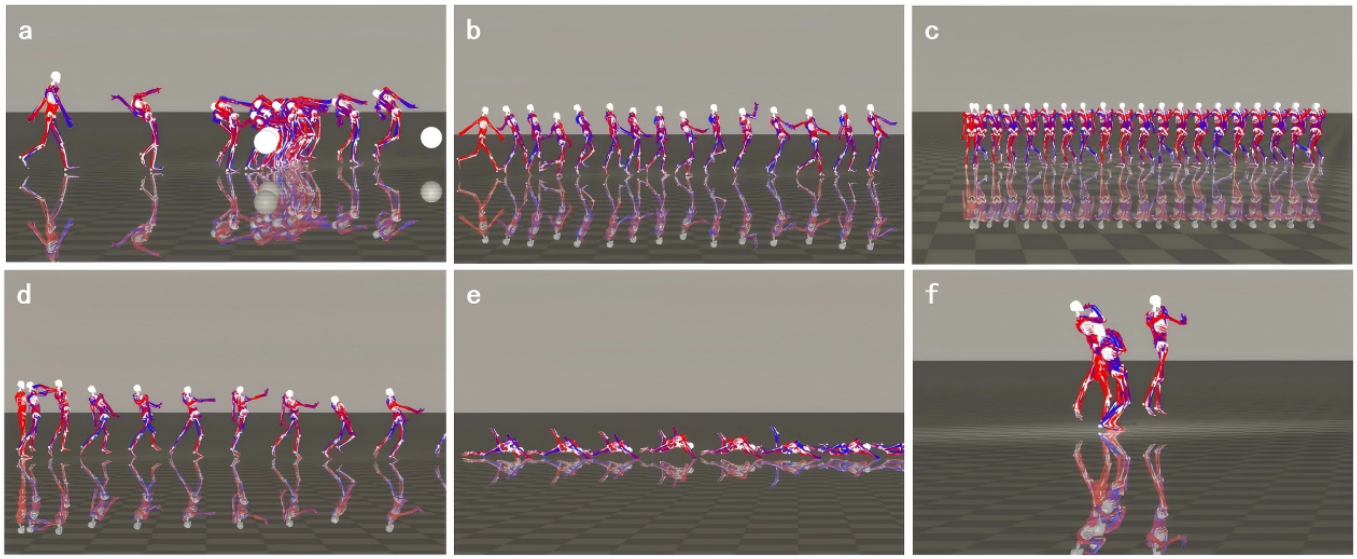

Fig. 12: **Visualization of representative locomotion behaviors generated by the humanoid robot.** Each subfigure shows a sequence of motion frames, with red and blue skeletons indicating the two legs and body segments over time. (a) **walk turn**: The robot moves while interacting with spherical obstacles. (b) **walk forward**: The robot demonstrates a stable periodic gait with alternating leg support. (c) **sidestep**: The robot demonstrates sustained rhythmic whole-body coordination. (d) **run**: The robot maintains balance while moving on a wheeled board. (e) **crawl**: The robot performs low-posture locomotion with the body close to the floor. (f) **Jump**: The robot performs a dynamic aerial transition while maintaining body balance.

for muscle effort and physiological efficiency of the generated motion. Since the human central nervous system tends to optimize energy consumption, this metric helps filter out poor-quality policies that maintain balance by relying on excessive muscle activation or co-contraction [78, 60]. Formally, the mean squared muscle effort is computed as:

$$E_{\text{muscle}} = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{M}\sum_{m=1}^{M} a_{m,t}^2\right), \tag{6}$$

where $a_{m,t} \in [0,1]$ denotes the actuator activation of the m-th muscle at time step t, T denotes the total number of simulation steps in each episode, and M denotes the total number of muscles. A lower E_{muscle} indicates that the policy accomplishes the task with better physiological efficiency and less unnecessary muscle effort.

4) Joint Smoothness: Joint smoothness is evaluated by analyzing high-frequency fluctuations in joint motion. Prior work has used smoothness-related kinematic measures to characterize coordinated and biologically plausible movement [28, 5]. Lower high-frequency jerk indicates smoother motor behavior with fewer abrupt joint-level changes, which is consistent with the low-pass characteristics of biological motor control [28, 5].

In our implementation, smoothness is computed from the joint angular velocity signal returned by the physics engine, denoted as $v_{j,t}$ for the j-th joint at time step t. Since our implementation applies a second-order finite difference to joint velocity, the resulting finite-difference quantity is proportional to the third-order derivative of joint angle, commonly referred to as jerk:

$$\begin{aligned} \mathrm{D}^{(2)} v_{j,t} &= \frac{v_{j,t} - 2v_{j,t-1} + v_{j,t-2}}{\Delta t^2} \\ &\approx \frac{\mathrm{d}^2 v_j(t)}{\mathrm{d}t^2} = \frac{\mathrm{d}^3 \theta_j(t)}{\mathrm{d}t^3}. \end{aligned} \tag{7}$$

We therefore quantify joint-level high-frequency fluctuation using the mean squared finite-difference jerk over all joints and time steps:

$$J_{\text{sq}} = \frac{1}{(N-2)N_{\text{joint}}}\sum_{t=3}^{N}\sum_{j=1}^{N_{\text{joint}}}\left(\mathrm{D}^{(2)} v_{j,t}\right)^2, \tag{8}$$

where N_{joint} denotes the total number of joints, N denotes the total number of simulation steps in the episode, $v_{j,t}$ denotes the joint angular velocity of the j-th joint at time step t, Δt denotes the simulation time step, and $\mathrm{D}^{(2)} v_{j,t}$ denotes the second-order finite difference of joint angular velocity.

For visualization and task-level comparison, we report the logarithmic mean squared jerk:

$$\text{Smoothness}_{\text{jerk}} = \log_{10}\left(J_{\text{sq}} + \epsilon\right), \tag{9}$$

where ϵ is a small constant used for numerical stability. Lower values indicate smoother motion with fewer high-frequency joint fluctuations, whereas higher values indicate less smooth and more abrupt joint motion.

5) Peak Efficiency Steps: We introduce this metric to identify the training phase at which an algorithm achieves its peak evaluation performance. Peak Efficiency Steps pinpoints the number of environment interactions required to obtain

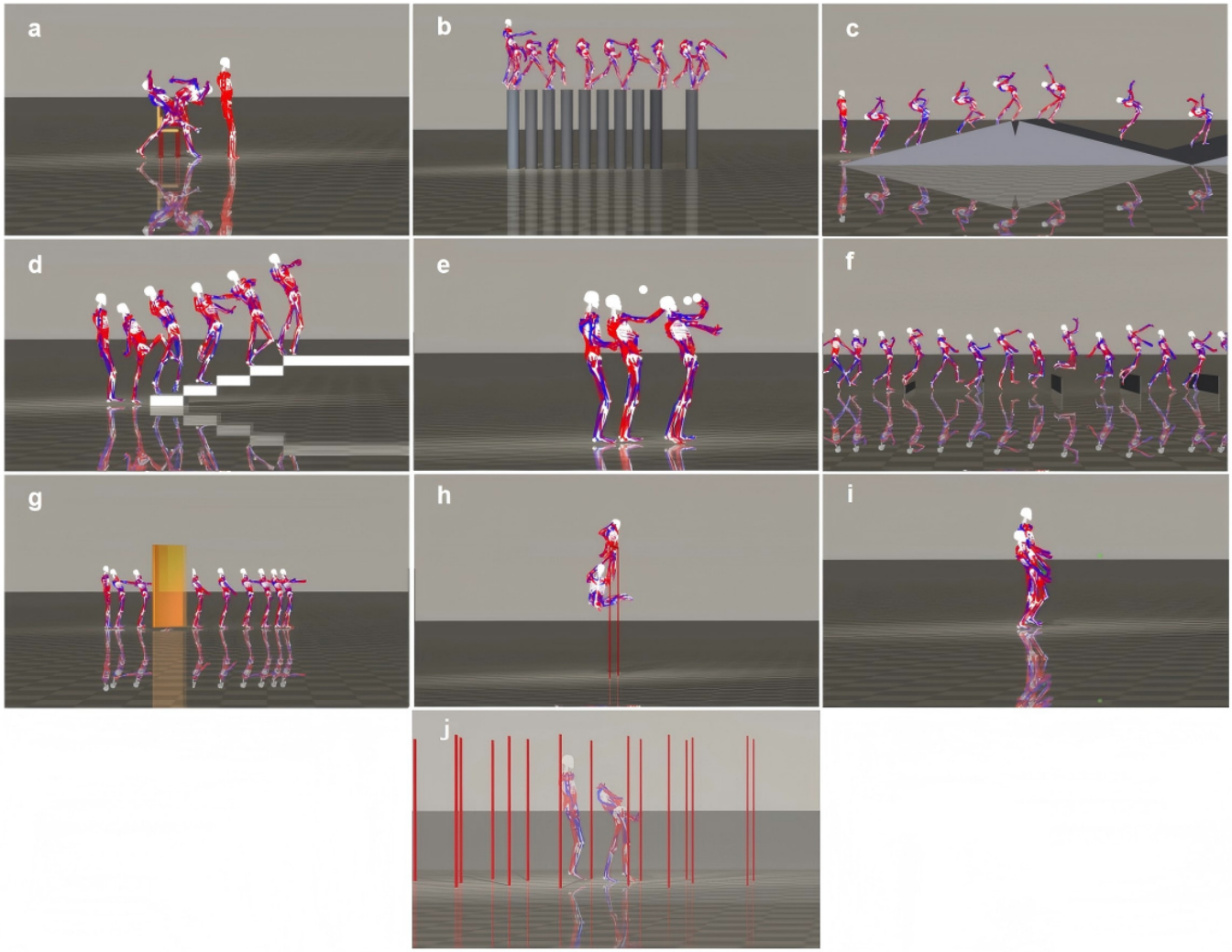

Fig. 13: **Illustration of ten representative terrain and object interaction tasks.** (a) **walkandsit**: The robot is required to approach a chair backward and sit down. Instead of using a conventional collision-avoidance penalty, the task rewards the robot for confidently utilizing downward inertia to complete the sitting motion and for comfortably leaning back against the chair after sitting, thereby reproducing human-like sitting biomechanics. (b) **stepstones**: A magnetic attraction reward is used to guide the foot toward the target landing positions, while strict penalties on trunk lateral bending, yaw deviation, and excessive knee flexion encourage an upright and stable traversal posture. (c) **slide**: The robot is required to adaptively lean forward to counteract inertia during slope transitions. A split-prevention penalty is introduced to terminate the episode once the lateral separation of the legs exceeds the biomechanical limit. (d) **stairs**: The robot is required to climb stairs using look-ahead terrain perception from a terrain radar. When upcoming steps are detected, an active leg-lifting reward is activated to encourage anticipatory step planning. (e) **reach**: The robot is required to rapidly and explosively reach randomly reset spatial targets with its dominant hand while maintaining a slightly crouched defensive posture. Since the target height varies, the task strictly constrains knee height to prevent cheating strategies such as kneeling. (f) **hurdle**: The task enforces a complete aerial flight phase and introduces a strict anti-flamingo-jump penalty to prevent abnormal strategies in which the robot crosses the hurdle by curling and dragging a single leg. (g) **open door**: The robot is required to approach, push open, and pass through a hinged door. A global attraction field guides the hand toward the door handle, while fall-prevention penalties discourage shortcut behaviors such as headbutting the door or diving through it. (h) **chinup**: The robot is required to pull its head above the horizontal bar from a hanging posture. The task rewards the activation of major pulling muscle groups such as the latissimus dorsi and biceps, while imposing strict penalties on body swinging and leg-kipping, forcing the motion to be completed through pure muscular strength. (i) **catch**: The robot is required to coordinate both arms to catch a flying cube thrown toward the chest while maintaining an upright posture. The key objective is to track the distance between both hands and the target, and a large sparse reward is given when the object is successfully caught and stabilized in mid-air. (j) **polewalk**: The robot is required to scan surrounding poles using terrain radar and traverse the entire pole region without collision. The episode is immediately terminated once any body part touches a pole.

TABLE V: Reward Functions for All 22 Musculoskeletal Benchmark Tasks (Part 1: Stabilization Series). Note: Action regularization ($R_{\text{efficiency}}$) and minor limit penalties ($\sum$ penalties) are applied broadly across tasks and abbreviated where standard.

Task	Term	Formula	Weight
1. Stand	Pose preservation	$\exp(-\lVert q_{\text{joints}} - q_{0,\text{joints}}\rVert)$	10
	Absolute stillness	$\exp(-10.0\lVert \mathbf{v}_{\text{CoM}}\rVert) \cdot \exp(-0.1\lVert \dot{q}\rVert)$	15
	Feet planted	$\exp(-10.0(\lVert \mathbf{v}_{\text{foot,L}}\rVert + \lVert \mathbf{v}_{\text{foot,R}}\rVert))$	10
	Torso balance	$\exp(-5.0 \max(0, 1 - \mathbf{u}_{\text{torso}} \cdot \hat{\mathbf{z}}))$	10
	Survival	$\mathbb{I}_{\text{alive}}$	5
2. Powerlift	Hold stability	$\mathbb{I}(z_{\text{bar}} > 1.6\text{m} \wedge \lvert \dot{z}_{\text{bar}}\rvert < 0.1\text{m/s})$	100
	Height threshold	$\mathbb{I}(z_{\text{bar}} > 1.6\text{m})$	50
	Upright alignment	$\exp(-20.0(1 - \mathbf{u}_{\text{torso}} \cdot \hat{\mathbf{z}}))$	100
	Lumbar spine safety	$\exp(-15.0\lvert q_{\text{lumbar}} - (-0.4)\rvert)$	80
	Muscular effort	$\text{mean}(a)$	30
	Hip hinge posture	$\mathbb{I}(q_{\text{lumbar}} > 0.1)$	200
3. Singleleg	Balance (CoM)	$\exp(-15.0\lVert \mathbf{p}^{xy}_{\text{pelvis}} - \mathbf{p}^{xy}_{\text{foot,R}}\rVert^2)$	500
	Lifted foot clear	$\text{clip}(z_{\text{foot,L}}/0.3, 0, 1)$	300
	Absolute stillness	$\exp(-5.0\lVert \mathbf{v}_{\text{CoM}}\rVert^2)$	400
	Upright alignment	$\exp(-10.0(1 - \mathbf{u}_{\text{torso}} \cdot \hat{\mathbf{z}})^2)$	200
	Foot touch penalty	$\mathbb{I}(z_{\text{foot,L}} < 0.05\text{m})$	-800
	Lateral lean penalty	$\lvert \mathbf{u}_{\text{torso}} \cdot \hat{\mathbf{y}}\rvert$	-200
4. Sit	Chair approach	$2.0/(\lVert \Delta\mathbf{p}^{xy}_{\text{chair}}\rVert + 0.2)$	10
	3D Contact docking	$\exp(-\frac{\Delta x^2}{2\cdot 0.15^2}) \cdot \exp(-\frac{\Delta z^2}{2\cdot 0.15^2})$	30
	Lateral centering	$\exp(-\frac{\Delta y^2}{2\cdot 0.05^2})$	30
	Seated leg geometry	$\exp(-2.0\lvert \bar{q}_{\text{hip}} - 1.5\rvert - k_{\text{knee}}\lvert \bar{q}_{\text{knee}} - 1.6\rvert)$	40
	Feet tucked under	$\exp(-5.0 \sum_i \lvert (x_{\text{foot},i} - x_{\text{pelvis}}) - 0.2\rvert)$	30
	Soft landing	$\exp(-2.0\lvert \dot{z}_{\text{pelvis}}\rvert)$	5
5. Balance	Forward lean stance	$\exp(-10.0(\theta_{\text{pitch}} - (-0.15))^2)$ (gated)	15
	CoM over board	$\exp(-20.0\lVert \mathbf{p}^{xy}_{\text{CoM}} - \mathbf{p}^{xy}_{\text{board}}\rVert^2)$	15
	Board over ball	$\exp(-5.0\lVert \mathbf{p}^{xy}_{\text{board}} - \mathbf{p}^{xy}_{\text{ball}}\rVert)$	5
	Board level	$\exp(-5.0(1 - \mathbf{u}_{\text{board}} \cdot \hat{\mathbf{z}}))$	5
	Stand tall	$\exp(-10.0(1.30 - z_{\text{pelvis}})^2)$ if $z < 1.3$	4
6. Squat	Height tracking	$\exp(-30.0(z_{\text{pelvis}} - z_{\text{target}})^2)$	200
	Joint tracking	$\exp(-5.0(\bar{q} - q_{\text{target}})^2)$	300/200
	Vertical vel tracking	$\exp(-5.0(v_{z,\text{CoM}} - v_{z,\text{target}})^2)$	200
	CoM centered	$\exp(-15.0\lVert \mathbf{p}^{xy}_{\text{pelvis}} - \mathbf{p}^{xy}_{\text{mid_feet}}\rVert^2)$	300
	Kneeling penalty	$\mathbb{I}(\min(z_{\text{knee,L}}, z_{\text{knee,R}}) < 0.25\text{m})$	-1000

the highest expected evaluation return over the full training horizon. Formally, it is defined as the evaluated training step t that maximizes the expected evaluation return:

$$T_{\max} = \arg\max_{t \in \mathcal{T}} \mathbb{E}\left[R_t\right], \tag{10}$$

where $\mathcal{T}$ denotes the set of evaluated training steps throughout the learning process, and R_t denotes the cumulative reward obtained during evaluation at training step t. We also denote the corresponding maximum expected return as

$$R_{\max} = \max_{t \in \mathcal{T}} \mathbb{E}\left[R_t\right]. \tag{11}$$

A smaller $T_{\max}$, particularly when paired with a high $R_{\max}$, indicates that the algorithm can rapidly discover a high-performing behavioral policy, reflecting better sample efficiency and faster peak skill acquisition.

6) EMG Similarity: Finally, EMG similarity is evaluated by comparing simulated muscle activation envelopes with processed walking human EMG envelopes from Boo et al. [7], following the same EMG-envelope interpretation used in MuscleMimic [51]. We use only the walking EMG recordings from Boo et al. [7] for this metric. Since multiple muscle coordination strategies can generate similar joint kinematics, and accurate motion tracking does not necessarily imply human-like muscle recruitment, we treat the simulated muscle activation state as an activation-envelope proxy [51, 83, 60].

For each selected lower-limb muscle, simulated activations and human EMG recordings are processed in a gait-cycle-normalized form. Human EMG envelopes from Boo et al. [7] are rectified, normalized on a per-muscle basis, temporally aligned to the walking gait cycle, and resampled to a common normalized stride representation [51, 7]. Muscle activations from RL are segmented into gait cycles, interpolated to the same normalized time base, and averaged across cycles to obtain representative activation profiles. We then compare the

TABLE VI: Reward Functions for Musculoskeletal Benchmark Tasks (Part 2: Locomotion Series).

Task	Term	Formula	Weight
1. Walk	Velocity tracking	$\exp(-2.0\lVert\mathbf{v}_{\text{CoM}} - \mathbf{v}_{\text{target}}\rVert^2)$	10
	Kinematic imitation	$\exp(-2.0\lVert q_{\text{joints}} - q_{\text{ref,joints}}\rVert^2)$	3
	Upright / No hunch	$\exp(-10.0(1 - \mathbf{u}_{\text{torso}} \cdot \hat{\mathbf{z}}))$, $\exp(-20.0(\mathbf{u}_{\text{chest}} \cdot \hat{\mathbf{z}} + 0.1)^2)$	6/8
	Forward face	$\exp(-5.0(1 - \mathbf{u}_{\text{chest}} \cdot \hat{\mathbf{x}}))$	8
2. Crawl	Backwards velocity	$\text{clip}(-v_{x,\text{CoM}}/v_{\text{target}}, -0.2, 1.0)$	500
	Prone posture	$\exp(-5(\mathbf{u}_{\text{chest}} \cdot \hat{\mathbf{z}} + 1)^2) \cdot \ldots$	100
	Alternating stride	$\text{clip}(\lvert\Delta x_{\text{hands}}\rvert/0.3, 0, 1) \cdot \ldots$ (diag gait)	150
	Drift penalty	$2.0\lvert v_{y,\text{CoM}}\rvert + \lvert v_{x,\text{CoM}}\rvert\mathbb{I}(v_x > 0)$	-150
3. Run	High velocity	$\text{clip}(v_{x,\text{CoM}}, -0.5, 6.0)$, bonus if $v > 3.0$	200/100
	Asymmetric stride	$\text{clip}(d_{x,\text{feet}} + 2.0 \cdot d_{z,\text{feet}}, 0.0, 1.0)$	200
	Knee flexion (swing)	$\text{clip}((\max(q_{\text{knee}}) - 0.1)/1.0, 0, 1)$	200
	Scissor penalty	$\text{clip}((0.15 - d_{y,\text{feet}})/0.15, 0, 1)$	-300
	Kangaroo penalty	Penalize $d_{x,\text{feet}} < 0.2 \wedge d_{z,\text{feet}} < 0.1$	-300
4. Jump	Flight gradient	$\text{clip}(\frac{1}{2}(z_{\text{foot,L}} + z_{\text{foot,R}})/0.1, 0, 1)$	600
	Upward explosion	$\text{clip}(\dot{z}_{\text{pelvis}}, 0, 2.0)$	300
	Feet separation pen.	$d_{xy,\text{feet}}$	-200
	Desync jump pen.	$\lvert q_{\text{knee,L}} - q_{\text{knee,R}}\rvert + \lvert q_{\text{hip,L}} - q_{\text{hip,R}}\rvert$	-150
5. Walk turn	Targeted velocity	$\text{clip}(\mathbf{v}_{\text{CoM}} \cdot \mathbf{u}_{\text{target}}/3.0, -1.0, 1.0)$	200
	Waypoint approach	$\exp(-\lVert\Delta\mathbf{p}_{\text{target}}\rVert)$	50
	Target facing	$\text{clip}(\mathbf{u}_{\text{fwd}} \cdot \mathbf{u}_{\text{target}}, 0, 1)$	50
	Waypoint reached	$\mathbb{I}(\lVert\Delta\mathbf{p}_{\text{target}}\rVert < 0.5\text{m})$	500
	Wall collision	$\mathbb{I}_{\text{contact}}(\text{wall})$	-200
6. Sidestep	Lateral velocity	$\text{clip}(v_{y,\text{CoM}}/v_{\text{target}}, -0.2, 1.0)$	400
	Phase sync	$\exp(-15.0(e_L + e_R))$	200
	CoM mid-feet lock	$\lvert y_{\text{pelvis}} - \frac{1}{2}(y_{\text{foot,L}} + y_{\text{foot,R}})\rvert$	-500
	Rubber band splits	$\max(0, d_{xy,\text{feet}} - 0.3)$	-400

waveform similarity between the simulated activation envelope and the processed human EMG envelope using Pearson correlation:

$$r_m = \frac{\sum_{t=1}^{T} \left(a_{\text{sim},m}(t) - \bar{a}_{\text{sim},m}\right)\left(e_{\text{EMG},m}(t) - \bar{e}_{\text{EMG},m}\right)}{\sqrt{\sum_{t=1}^{T} \left(a_{\text{sim},m}(t) - \bar{a}_{\text{sim},m}\right)^2}} \times \frac{1}{\sqrt{\sum_{t=1}^{T} \left(e_{\text{EMG},m}(t) - \bar{e}_{\text{EMG},m}\right)^2}}, \tag{12}$$

where $a_{\text{sim},m}(t)$ denotes the gait-cycle-normalized simulated activation envelope of muscle m at normalized time t, $e_{\text{EMG},m}(t)$ denotes the corresponding processed walking EMG envelope from Boo et al. [7], and $\bar{a}_{\text{sim},m}$ and $\bar{e}_{\text{EMG},m}$ denote their temporal means. We report both per-muscle correlations and the average correlation across the selected muscles:

$$r_{\text{EMG}} = \frac{1}{M} \sum_{m=1}^{M} r_m, \tag{13}$$

where M is the number of muscles included in the comparison.

Higher r_m indicates stronger agreement in the timing and shape of the activation waveform. This metric should be interpreted as a measure of EMG-envelope similarity rather than direct proof of identical physiological control. High values indicate that the simulated policy recruits a muscle with a temporal pattern similar to human walking EMG, while lower values may reflect alternative but mechanically valid recruitment strategies caused by muscle redundancy. Following MuscleMimic [51], EMG similarity is therefore used as a physiological-plausibility diagnostic that complements task success and kinematic tracking, rather than as a unique ground-truth measure of internal control fidelity.

C. Detailed Descriptions of Reward-Based RL Baselines

In Section III, we categorized several standard and exploration-enhanced reinforcement learning methods under the reward-based RL methods. The detailed descriptions of these individual baselines are provided below:

- **PPO (Proximal Policy Optimization).** This [76] serves as the quintessential on-policy actor–critic baseline for continuous-control tasks. By employing a clipped surrogate objective function, PPO ensures conservative policy updates, preventing destructively large parameter shifts that frequently occur in highly nonlinear biomechanical simulations. While it offers excellent training stability and reliable convergence, its purely reward-driven, unstructured exploration often proves sample-inefficient when navigating the high-dimensional, highly redundant muscle excitation spaces inherent to full-body musculoskeletal humanoids.
- **SAC (Soft Actor-Critic).** This [33] is utilized as the primary off-policy baseline, distinguished by its maximum-

TABLE VII: Reward Functions for Musculoskeletal Benchmark Tasks (Part 3: Interaction with Environment Series).

Task	Term	Formula	Weight
1. Stairs	Fwd trajectory	$\text{clip}(v_{x,\text{CoM}}, -0.2, 2.0)$	50
	Upward trajectory	$\text{clip}(v_{z,\text{CoM}}, -0.5, 3.0)$	30
	Vertical stable	$\exp(-5.0(0.98 - \mathbf{u}_{\text{torso}} \cdot \hat{\mathbf{z}})^2)$	20
	Lateral stable	$\exp(-5.0 y_{\text{pelvis}}^2)$	20
	Anticipatory lift	$\mathbb{I}(\text{radar}) \cdot \max(\dot{z}_{\text{swing}}, 0)$	40
2. Hurdle	Forward velocity	$\text{clip}(v_{x,\text{CoM}}, -0.1, 4.0)$	300
	Clearance (Flight)	$\mathbb{I}(\max(z_{\text{feet}}) > h_{\text{target}})$	100
	Sparse clear bonus	$\mathbb{I}(\Delta x < 0 \wedge \lvert y_{\text{pelvis}} \rvert < 0.5\text{m})$	3000
	Crane hopping pen.	$\text{clip}((\lvert z_{\text{foot,L}} - z_{\text{foot,R}} \rvert - 0.2)/0.5, 0, 1)$	-150
3. Step stones	Target stone progress	Bonus when crossing stone X-coord	800
	Magnetic reaching	$\exp(-2.0 \cdot d_{\text{lead}}) \cdot \mathbb{I}(v_{x,\text{CoM}} > 0.1)$	150
	Stand tall	$\text{clip}((\Delta z_{\text{leg}} - 0.6)/0.3, 0, 1)$	50
	Overbend knee pen.	$\max(q_{\text{knee}} - 1.0, 0)$	-50
4. Slide	Pitch balance	$R_{\text{gate}} \cdot \exp(-5.0 \cdot \theta_{\text{pitch}}^2)$	2
	Stand tall	$R_{\text{gate}} \cdot \exp(-10.0(0.95 - \Delta z_{\text{leg}})^2)$	2
	Slope forward lean	$\text{clip}(-\mathbf{u}_{\text{chest}} \cdot \hat{\mathbf{z}}/0.3, 0, 1)$	20
	Backward lean pen.	$\max(\mathbf{u}_{\text{chest}} \cdot \hat{\mathbf{z}}, 0)$	-100
5. Open door	Hand attractor field	$\exp(-1.5 \lVert \Delta\mathbf{p}_{\text{handle}} \rVert)$	400
	Open hinge	$\text{clip}(\theta_{\text{door}}/1.2, 0, 1) \cdot \text{prox_gate}$	500
	Pass threshold	Sparse bonuses for $x > 1.0, 2.0$	2000
	Dive/Split penalties	Penalize pelvis lead or wide feet	$-300/-400$
6. Reach	Dense reach dist	$\exp(-4.0 \lVert \Delta\mathbf{p}_{\text{reach}} \rVert)$	50
	Target velocity proj	$\text{clip}(\mathbf{v}_{\text{hand}} \cdot \mathbf{u}_{\text{reach}}, 0, 2.0)$	20
	Target touch bonus	$\mathbb{I}(\lVert \Delta\mathbf{p}_{\text{reach}} \rVert < 0.15\text{m})$	200
7. Walk & sit	Backwards precision	$\exp(-\lVert \Delta\mathbf{p}_{\text{chair}}^{xy} \rVert^2/0.005)$	100
	Fast descent bonus	$\mathbb{I}(\dot{z}_{\text{pelvis}} < -1.5\text{m/s})$	50
	Relaxed recline	$\mathbb{I}(\mathbf{u}_{\text{torso}} \cdot \hat{\mathbf{z}} < 0.8)$ (once seated)	200
	90-degree pose	$\exp(-5(\bar{q}_{\text{hip}} - 1.5)^2) + \ldots$	100
8. Chin up	Spatial gradient	$\exp(-5.0 \max(0, h_{\text{bar}} - z_{\text{head}}))$	50
	Muscle force filter	$\text{mean}(F_{\text{pull_muscles}})$	5
	Explosive accel	$\max(0, \ddot{z}_{\text{pelvis}})$	40
	Anti-kipping wiggle	$\lVert \dot{q}_{\text{legs}} \rVert$	-20
9. Catch	Hand-object reach	$\exp(-3.0 \cdot d_{\text{catch}})$	100
	Interception bonus	$\mathbb{I}(d_{\text{catch}} < 0.15\text{m} \wedge z_{\text{cube}} > 0.5\text{m})$	500
	Drop penalty	$\mathbb{I}(z_{\text{cube}} < 0.1\text{m})$	-200
10. Pole walk	Target feet phases	$\exp(-10.0(\Delta x_{\text{foot}} \pm 0.25 \sin\phi)^2)$	400
	Dynamic radar gap	E.g., $\text{clip}(2.0 \cdot v_{y,\text{CoM}}, 0, 1)$ if left gap	800
	Fatal collision	$\mathbb{I}_{\text{fatal_contact}}$ (head/torso to pole)	-1500

TABLE VIII: Task Settings, Specific Observations, and Termination Criteria (Part 1: Stabilization Series). *Base Observation (O_{base}): All tasks fundamentally include proprioceptive states $(q, \dot{q}, \mathbf{v}_{CoM})$ and muscle states (length, velocity, force, activation). Only task-specific exteroceptive or goal variables are listed below.*

Task	Specific Obs (O_{task})	Initialization (Brief)	Termination Criteria
1. Stand	Torso quat, feet height	Default neutral, zero vel.	$z_{\text{CoM}} < 0.8$ or tilt > 0.8
2. Powerlift	Dumbbell $\mathbf{p}_{\text{bar}}$, $\Delta\mathbf{p}_{\text{bar}}$, tilt	High CoM, pre-loaded lumbar	$z_{\text{pelvis}} < 0.7$ or $q_{\text{lumbar}} > 0.5$
3. Singleleg	Torso quat, feet/CoM z	Flamingo pose, noise	$z_{\text{pelvis}} < 0.6$, dropped foot
4. Sit	Chair relative $\Delta\mathbf{p}_{\text{chair}}$	In front of chair, hips pre-flexed	$z_{\text{pelvis}} < 0.25$, invalid contact
5. Balance	Board $\mathbf{p}, \mathbf{q}$, CoM/Ball $\Delta\mathbf{p}$	Drop $z = 1.28$m, guard stance	$z_{\text{pelvis}} < 0.70$, pitch fall
6. Squat	Phase ϕ, target z_{target}	Squat start pose, barbell hold	$z_{\text{pelvis}} < 0.35$, kneeling

TABLE IX: Task Settings, Specific Observations, and Termination Criteria (Part 2: Locomotion Series).

Task	**Specific Obs (O_{task})**	**Initialization (Brief)**	**Termination Criteria**
1. Walk	Phase ϕ, error $\Delta q = q - q_{\text{ref}}$	Sampled from synthetic walk	$z_{\text{CoM}} < 0.5$ or tilt > 0.8
2. Crawl	Torso quat, feet height	Prone quadruped, ground $z = 0.25$	Fwd move, twist, flip
3. Run	Torso angle, feet height	Standard standing	$z_{\text{pelvis}} < 0.45$, fall backward
4. Jump	Torso angle, $d_{xy,\text{feet}}$	Micro-squat, pre-flexed	$z_{\text{pelvis}} < 0.45$, wide split
5. Walk turn	Target $\Delta\mathbf{p}$, ID, facing a	Stand, 20-step novice protection	$z_{\text{pelvis}} < 0.55$, $z_{\text{head}} < 0.8$
6. Sidestep	Phase ϕ, lat dist Δy_{feet}	Crab guard stance, hips wide	$z_{\text{pelvis}} < 0.45$, split > 0.5m

TABLE X: Task Settings, Specific Observations, and Termination Criteria (Part 3: Interaction with Environment Series).

Task	**Specific Obs (O_{task})**	**Initialization (Brief)**	**Termination Criteria**
1. Stairs	Phase ϕ, radar $\mathbf{z}_{\text{terrain}} \in \mathbb{R}^9$	Walk posture, initial fwd vel	$z_{\text{pelvis}} < 0.75$, lat drift
2. Hurdle	Phase ϕ, nearest $\Delta\mathbf{h}$	Standing, scan dynamic hurdles	$z_{\text{pelvis}} < 0.55$, heavy yaw
3. Step stones	Phase ϕ, stones $\Delta\mathbf{p}_{\text{stones}}$	Elevated platform $z = 3.05$m	Fall from height, $z_{\text{pelvis}} < 2.5$
4. Slide	Phase ϕ, normal $\mathbf{n}_{\text{slope}}$	Elevated over wave terrain	Backwards fall, split > 1.2m
5. Open door	Handle $\Delta\mathbf{p}$, door θ, target	Stand front of closed door	$z_{\text{pelvis}} < 0.65$, head < 1.1
6. Reach	Phase ϕ, reach $\Delta\mathbf{p}_{\text{reach}}$	Quarter-squat, hands raised	Knee drop < 0.15m
7. Walk & sit	Chair $\Delta\mathbf{p}$, phase $\mathbb{I}_{\text{seated}}$	Face away, inject backward vel	Ground contact (head/knees)
8. Chin up	Phase (0), bar dist Δz_{chin}	Hanging, flexed shoulders	Fall off bar ($z_{\text{pelvis}} < 0.5$)
9. Catch	Cube $\mathbf{p}, \mathbf{v}$, dist d_{catch}	Anticipatory basket, cube spawn	Drop cube ($z_{\text{cube}} < 0.05$)
10. Pole walk	Phase ϕ, radar $\mathbf{z}_{\text{radar}} \in \mathbb{R}^7$	Elevated 1.05m, high core gain	Fatal upper-body collision

TABLE XI: Benchmark Performance and Robustness Evaluation: Stabilization

Task	**Algorithm**	**Energy Consump.**	**Robustness** Action					Observation					Dynamic					**Max Steps**	**Success Rate**	$\log_{10}$**smooth (rad/s^3)**
			0	0.05	0.1	0.15	0.2	0	0.02	0.05	0.08	0.10	0	0.05	0.1	0.15	0.2			
stand	DepRL	0.4488	94	94	90	96	98	100	12	0	0	0	98	48	28	12	4	3.74	86	3.152
	SAC	0.0484	100	100	100	100	96	100	0	0	0	0	100	100	100	100	100	7.13	100	1.875
	PPO	0.3691	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.97	0	5.479
	DynSyn-SAC	0.2859	52	42	56	46	36	46	46	22	30	14	40	32	40	30	30	5.44	52	2.104
powerlift	DepRL	0.3356	68	68	64	56	52	62	62	38	4	0	66	70	52	42	38	4.14	38	2.631
	SAC	0.2106	98	94	94	88	80	96	0	0	0	0	98	94	88	64	2	6.78	90	2.825
	PPO	0.3955	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.76	0	5.042
squat	DepRL	0.3781	44	46	38	32	30	24	56	34	62	26	46	38	12	20	10	14.02	42	2.984
	SAC	0.1184	0	0	0	4	0	0	0	0	0	0	4	0	0	0	0	4.79	0	1.708
	PPO	0.3455	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2.01	0	5.390
sit	DepRL	0.3327	96	100	100	98	92	92	100	100	100	94	100	92	96	100	90	4.88	96	2.872
	SAC	0.1164	80	80	74	70	64	84	0	0	0	0	78	100	76	70	52	7.39	64	3.204
	PPO	0.4517	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2.37	0	5.350
	Dyn	0.3302	80	72	70	68	60	80	80	84	82	78	92	84	78	60	26	9.27	82	3.145
singlestand	DepRL	0.3977	14	16	28	18	10	24	30	8	0	0	12	28	6	8	12	2.10	14	2.854
	SAC	0.0849	0	0	0	0	0	3.3	0	0	0	0	0	0	0	0	0	4.33	0	1.954
	PPO	0.3981	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.84	0	5.436
balance	DepRL	0.3775	60	40	26	20	14	56	0	0	0	0	46	12	36	22	18	8.86	26	2.822
	SAC	0.1459	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8.11	0	2.547
	PPO	0.3859	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.25	0	4.861

TABLE XII: Benchmark Performance and Robustness Evaluation: Locomotion

Task	Algorithm	Energy Consump.	Robustness(%) Action					Observation					Dynamic					Max Steps($\times 10^7$)	Success Rate(%)	$\log_{10}$smooth (rad/s^3)
			0	0.05	0.1	0.15	0.2	0	0.02	0.05	0.08	0.10	0	0.05	0.1	0.15	0.2			
walk forward	DepRL	0.4206	28	28	32	32	44	34	32	46	24	30	40	20	12	8	6	3.32	64	3.151
	SAC	0.1427	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9.88	0	3.262
	PPO	0.3940	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2.42	0	5.558
run	DepRL	0.3354	30	22	32	36	42	18	18	30	32	26	32	16	12	8	2	6.72	18	3.343
	SAC	0.0940	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2.13	0	2.854
	PPO	0.3696	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2.16	0	5.524
walk turn	DepRL	0.3385	46	60	62	66	68	34	40	58	40	26	48	46	40	36	36	6.38	42	3.057
	SAC	0.1571	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.99	0	2.610
	PPO	0.3747	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.05	0	5.331
jump	DepRL	0.3616	8	24	20	12	10	18	18	8	4	0	14	10	12	6	0	6.89	10	3.188
	SAC	0.1342	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3.92	0	2.853
	PPO	0.3880	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2.07	0	5.640
sidestep	DepRL	0.3888	40	38	62	56	32	34	12	2	0	0	24	22	18	10	6	4.38	28	3.349
	SAC	0.1045	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.000032	0	3.166
	PPO	0.3729	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4.92	0	5.408
	DynSyn-SAC	0.3630	74	66	54	42	30	74	78	74	78	54	64	72	60	40	40	5.58	76	3.223
crawl	DepRL	0.3540	100	100	100	100	98	96	100	100	100	100	100	100	98	100	100	5.12	86	3.158
	SAC	0.0338	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2.87	0	2.516
	PPO	0.4036	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.34	0	5.674

entropy reinforcement learning framework. By jointly maximizing expected return and policy entropy, SAC explicitly encourages broad and robust exploration of the action space. This stochastic exploration strategy is particularly advantageous for musculoskeletal control: the entropy regularization prevents premature convergence to sub-optimal local minima—a common pathology in muscle-driven systems where multiple distinct muscle activation patterns can yield identical kinematic joint torques. Furthermore, its off-policy nature and replay buffer significantly improve sample efficiency over on-policy methods.

- **DepRL.** This [77] represents a specialized, exploration-enhanced reinforcement learning baseline tailored for musculoskeletal systems. Traditional independent Gaussian action noise often fails to activate muscles in a biologically coordinated manner, typically resulting in conflicting co-contractions and "stiff" joint behaviors. DepRL addresses this by augmenting the RL policy with Differential Extrinsic Plasticity (DEP)—an unsupervised, self-organizing exploration mechanism. Instead of injecting blind stochastic noise, the DEP controller dynamically updates a sensorimotor mapping matrix based on the temporal correlations between recent motor commands and the derivatives of proprioceptive feedback. This generates highly structured, mechanically grounded exploratory actions that intrinsically leverage the agent's physical embodiment. By superimposing these synergistic exploration patterns onto the RL policy outputs, DepRL effectively mitigates stiff co-contractions and significantly accelerates the discovery of functional locomotion gaits in highly overactuated models.
- **DynSyn-SAC.** This [36] is a state-of-the-art action-representation algorithm designed specifically for efficient learning and control in overactuated embodied systems. Built upon the maximum-entropy SAC framework, DynSyn-SAC introduces a dynamical synergistic structure directly into the network architecture. Specifically, it groups functionally related actuators and processes state observations into a lower-dimensional latent representation. A dedicated *DynSyn Layer* then maps this latent output to full-body muscle activations using group-shared synergistic weights. To ensure stable exploration during the early phases of learning, the algorithm employs a linear scheduling mechanism that dynamically bounds and scales the amplitude of these synergistic weights over the course of training. This hierarchical approach allows the policy to effectively predict group-level synergistic actions alongside state-dependent, muscle-level fine-tuning, thereby significantly accelerating sample efficiency in high-dimensional musculoskeletal models.

D. Detailed Reward Curves for Reward-Based RL Baselines

In this appendix, we present the detailed training results of the reward-based RL baselines. The reward curves for all 22 tasks are shown in Figure 14.

E. Detailed Muscle Activation Analysis: 416 vs. 700 Muscle Models

In the main text, we discussed how higher anatomical resolution enables more physiologically accurate compensatory strategies during the **powerlift** task. To provide deeper insight into these internal control differences, we present the time-step resolved muscle activation heatmaps for both the 416-muscle (Fig. 15) and the highly detailed 700-muscle (Fig. 16) models.

TABLE XIII: Benchmark Performance and Robustness Evaluation: Interaction with environment

Task	Algorithm	Energy Consump.	Robustness(%) Action					Observation					Dynamic					Max Steps($\times 10^7$)	Success Rate(%)	$\log_{10}$smooth (rad/s^3)
			0	0.05	0.1	0.15	0.2	0	0.02	0.05	0.08	0.10	0	0.05	0.1	0.15	0.2			
stair	DepRL	0.3499	52	54	38	42	40	34	38	2	0	0	32	40	24	14	4	1.42	36	2.982
	SAC	0.1993	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6.98	0	2.459
	PPO	0.3854	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.57	0	5.303
	DynSyn-SAC	0.2759	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.92	0	2.854
catch	DepRL	0.3247	80	80	76	76	84	62	48	80	72	46	78	40	42	38	56	7.23	28	2.919
	SAC	0.1384	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.37	0	2.899
	PPO	0.3964	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3.12	0	5.300
hurdle	DepRL	0.3715	48	48	56	50	40	64	44	48	28	0	62	44	40	32	20	1.26	50	3.212
	SAC	0.1852	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4.12	0	2.816
	PPO	0.3668	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2.46	0	5.398
slide	DepRL	0.3330	100	22	22	16	12	100	16	4	4	0	100	8	14	4	6	11.32	100	3.106
	SAC	0.0925	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4.66	0	2.752
	PPO	0.3934	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4.13	0	5.345
	DynSyn-SAC	0.2649	0	2	2	0	0	0	0	0	0	0	0	2	2	0	0	3.83	0	2.864
steppingstones	DepRL	0.3606	42	46	44	40	38	46	42	10	20	2	54	38	24	18	18	8.32	58	2.985
	SAC	0.1647	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7.33	0	2.140
	PPO	0.4020	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.27	0	5.211
polewalk	DepRL	0.4080	22	38	16	40	20	42	26	6	0	0	30	26	20	16	16	5.14	12	2.876
	SAC	0.1504	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6.10	0	1.724
	PPO	0.4274	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2.87	0	5.377
reach	DepRL	0.3410	84	86	84	94	82	94	82	90	62	34	92	88	82	82	74	9.34	84	2.975
	SAC	0.1705	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.85	0	2.677
	PPO	0.3965	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2.87	0	5.152
	DynSyn-SAC	0.2556	4	14	0	4	4	2	6	6	2	4	10	8	6	4	0	2.10	0	2.840
walkandsit	DepRL	0.3648	30	26	28	14	10	28	36	32	18	10	22	28	22	14	12	2.68	16	3.365
	SAC	0.1180	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6.54	0	2.410
	PPO	0.3951	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3.03	0	5.636
chinup	DepRL	0.3998	50	59	50	62	54	46	72	64	62	56	70	62	64	60	52	5.46	32	2.862
	SAC	0.4618	64	86	76	66	58	64	50	74	72	66	70	70	68	64	60	1.37	28	2.703
	PPO	0.3620	62	74	72	66	60	64	66	82	80	72	74	66	56	60	76	1.97	74	4.710
open door	DepRL	0.3304	96	100	100	100	98	100	94	74	62	42	100	96	96	78	66	4.71	98	3.030
	SAC	0.0962	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.97	0	2.836
	PPO	0.4112	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.77	0	5.327

Biomechanical Analysis. The heatmaps reveal a stark contrast in load-distribution strategies dictated by anatomical fidelity. In the 416-muscle model (Fig. 15), the agent is forced into an extreme compensatory state. Lacking the physiological infrastructure to properly brace the cervical and upper thoracic spine, it maxes out nearly all distal arm and superficial shoulder muscles (indicated by the overwhelming solid yellow patterns, representing $\sim$1.0 activation) simply to maintain the overhead lock.

Conversely, expanding to the 700-muscle model (Fig. 16) unlocks a significantly more human-like kinetic chain. A direct comparison of the shared upper-limb musculature (top section) reveals that the activation of these distal muscles is markedly reduced (showing more dynamic purple/green resting states). The massive mechanical stress on the shoulder girdle has been effectively offloaded. As demonstrated in the bottom section of Fig. 16 ("NECK & SPINE"), this stress is seamlessly transferred to the heavily recruited 700-exclusive axial stabilizers (e.g., the *multifidus*, *splenius*, and deep cervical erectors). By actively recruiting these deep neck and spine muscles, the 700-muscle agent optimally braces its core to safely support the heavy barbell. This comparison visually proves that under high-fidelity anatomical constraints, reward-driven optimization naturally converges on proper clinical force-transfer mechanics rather than relying on unrealistic distal compensations.

F. Agentic Reward Tuning

In our framework, we introduce *Agentic Reward Tuning*, a training mechanism that autonomously adjusts reward weights during reinforcement learning according to the agent's observed performance. Instead of relying on a fully fixed reward design, the reward composition is periodically refined by an external reasoning module based on recent gait quality, balance stability, and task-progress statistics.

Formally, let $\mathbf{r}_t$ denote the vector of sub-rewards at timestep t, and let $\mathbf{w}_t$ denote the corresponding reward weights. The dense reward is computed as:

$$R_t = \sum_i w_{i,t} r_{i,t}. \tag{14}$$

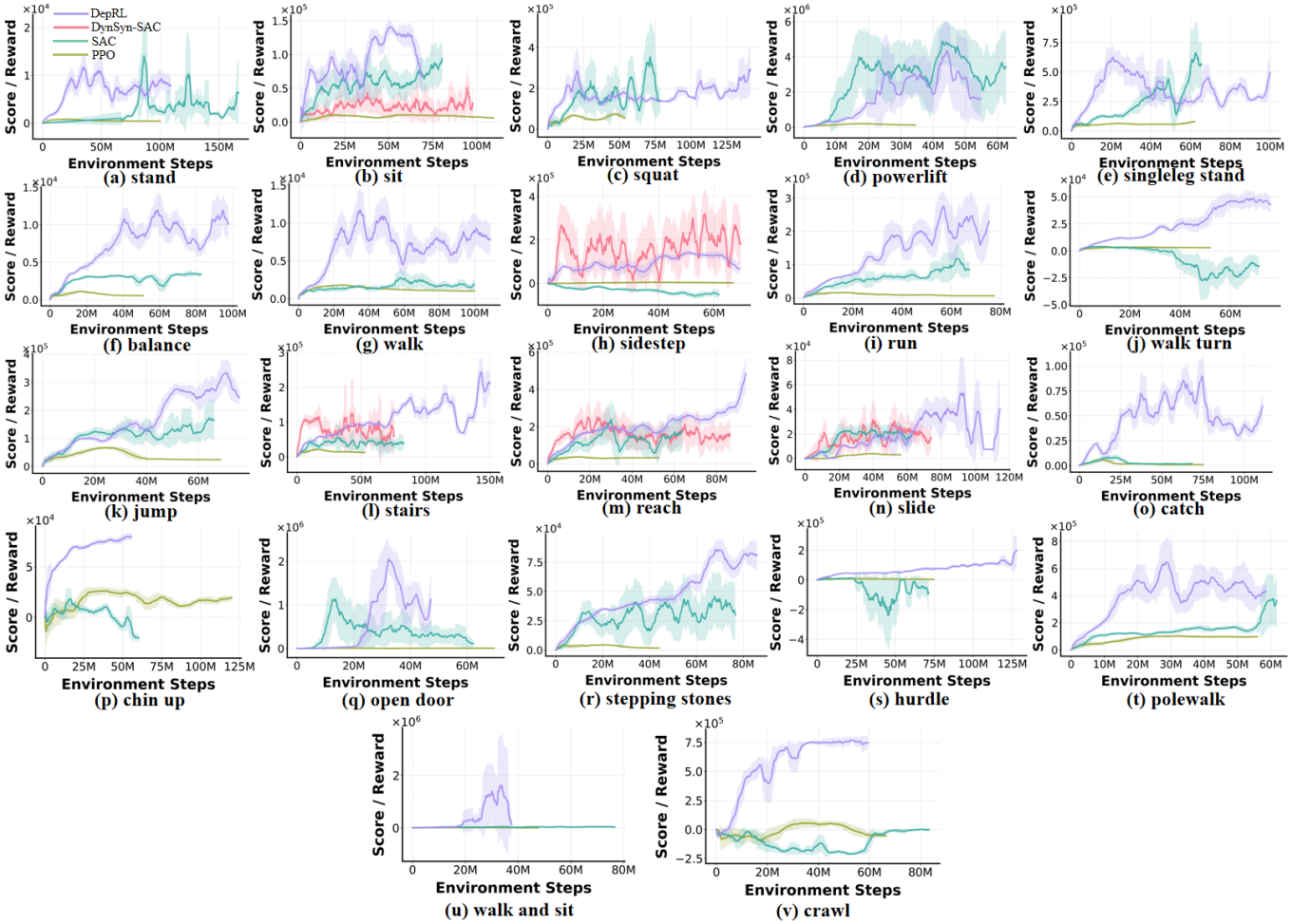

Fig. 14: **Learning curves for all 22 tasks.** This figure illustrates the detailed reward trajectories of the pure reinforcement learning baselines across the complete suite of 22 tasks during training.

During training, the reward weights are not treated as static hyperparameters. Every K environment steps, the system aggregates recent performance statistics:

$$\mathbf{s}_{\text{stats}} = \text{Aggregate}\left(\{\mathbf{o}_\tau, \mathbf{r}_\tau\}_{\tau=t-K}^{t}\right), \tag{15}$$

where $\mathbf{s}_{\text{stats}}$ contains task-relevant indicators such as forward velocity, torso pitch, posture stability, and tracking quality.

The reasoning module then updates the reward weights according to the current reward configuration and the collected statistics:

$$\mathbf{w}_{t+1} = \text{Reasoner}\left(\mathbf{w}_t, \mathbf{s}_{\text{stats}}\right), \tag{16}$$

with the update magnitude constrained to avoid destabilizing policy optimization:

$$\left\|\mathbf{w}_{t+1} - \mathbf{w}_t\right\|_\infty \leq \epsilon. \tag{17}$$

For example, if the agent exhibits insufficient forward speed, the tuning module increases the velocity-tracking weight. If the agent develops a hunched posture, the weights of uprightness and anti-hunching rewards are increased. In this way, Agentic Reward Tuning acts as an adaptive reward-shaping mechanism that continuously emphasizes the currently deficient aspects of the learned behavior.

In implementation, the tuning process is triggered periodically during environment rollout. Recent statistics, such as average forward speed and torso pitch, are cached over a fixed interval. A reasoning module analyzes these statistics and returns an updated reward-weight dictionary. To support parallel training, the updated weights are written to a shared file and synchronized across environments. A file-lock mechanism is used to prevent multiple workers from simultaneously calling the reasoning module, and fallback exception handling ensures that training continues even when the external call fails.

Overall, Agentic Reward Tuning provides a closed-loop reward refinement mechanism. It combines reinforcement learning with high-level performance diagnosis, allowing the policy to evolve under a dynamically adjusted reward landscape. This improves the practicality of reward engineering for complex musculoskeletal humanoid tasks, where manually balancing imitation, locomotion, posture, and physiological regularization terms is difficult.

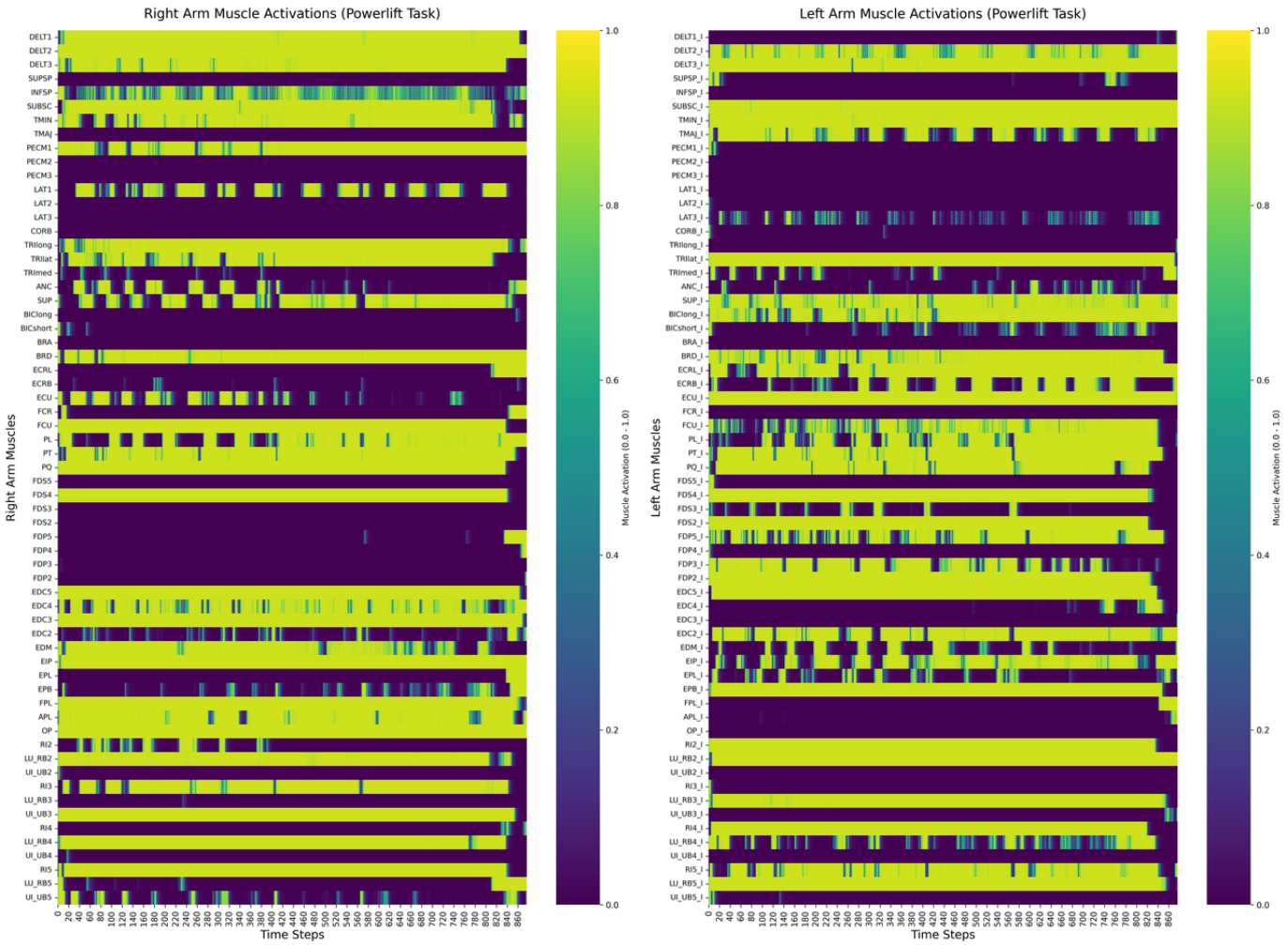

Fig. 15: **Upper limb muscle activation heatmap for the 416-muscle model (Powerlift).** The heatmap displays the intense, sustained activations of the available arm and shoulder musculature required to stabilize the barbell overhead. Note that this model lacks the detailed deep axial stabilizers of the neck and spine.

G. Latent-Action RL

The *Atent-Action RL* framework integrates expert-informed action filtering with reinforcement learning to guide musculoskeletal humanoid control. In the hard mode used in this work, the policy is not allowed to directly execute arbitrary high-dimensional muscle actions. Instead, each policy action is first projected through a pretrained anatomical encoder-decoder module, and the final executable action is constrained to remain within a bounded neighborhood of the expert-informed action. This design reduces unstable exploration in the high-dimensional muscle action space while preserving the adaptability of reinforcement learning.

1. Expert Teacher Pretraining: An expert teacher model is first trained using behavior cloning on high-quality demonstration data. The model adopts an encoder-decoder architecture:

- **Encoder:** maps the expert muscle action a_t to a low-dimensional latent intent z_t, which captures the underlying control intention.
- **Decoder (Full Anatomical Transformer):** reconstructs the muscle action conditioned on z_t and the current musculoskeletal state, including muscle priors, muscle states, and joint moment information.

Formally, for a batch of expert actions A and associated anatomical features (P, S, M), the reconstructed action is given by:

$$\hat{A} = \text{Decoder}(\text{Encoder}(A), P, S, M), \tag{18}$$

where P denotes muscle-level physical priors, S denotes muscle state features such as muscle length, velocity, force, and activation, and M denotes the muscle-joint moment matrix. The behavior cloning objective is:

$$\mathcal{L}_{\text{BC}} = \frac{1}{N} \sum_i \left\| \hat{A}_i - \frac{A_i + 1}{2} \right\|^2. \tag{19}$$

Here, the transformation $(A_i + 1)/2$ maps the original normalized action range $[-1, 1]$ to the muscle activation range $[0, 1]$.

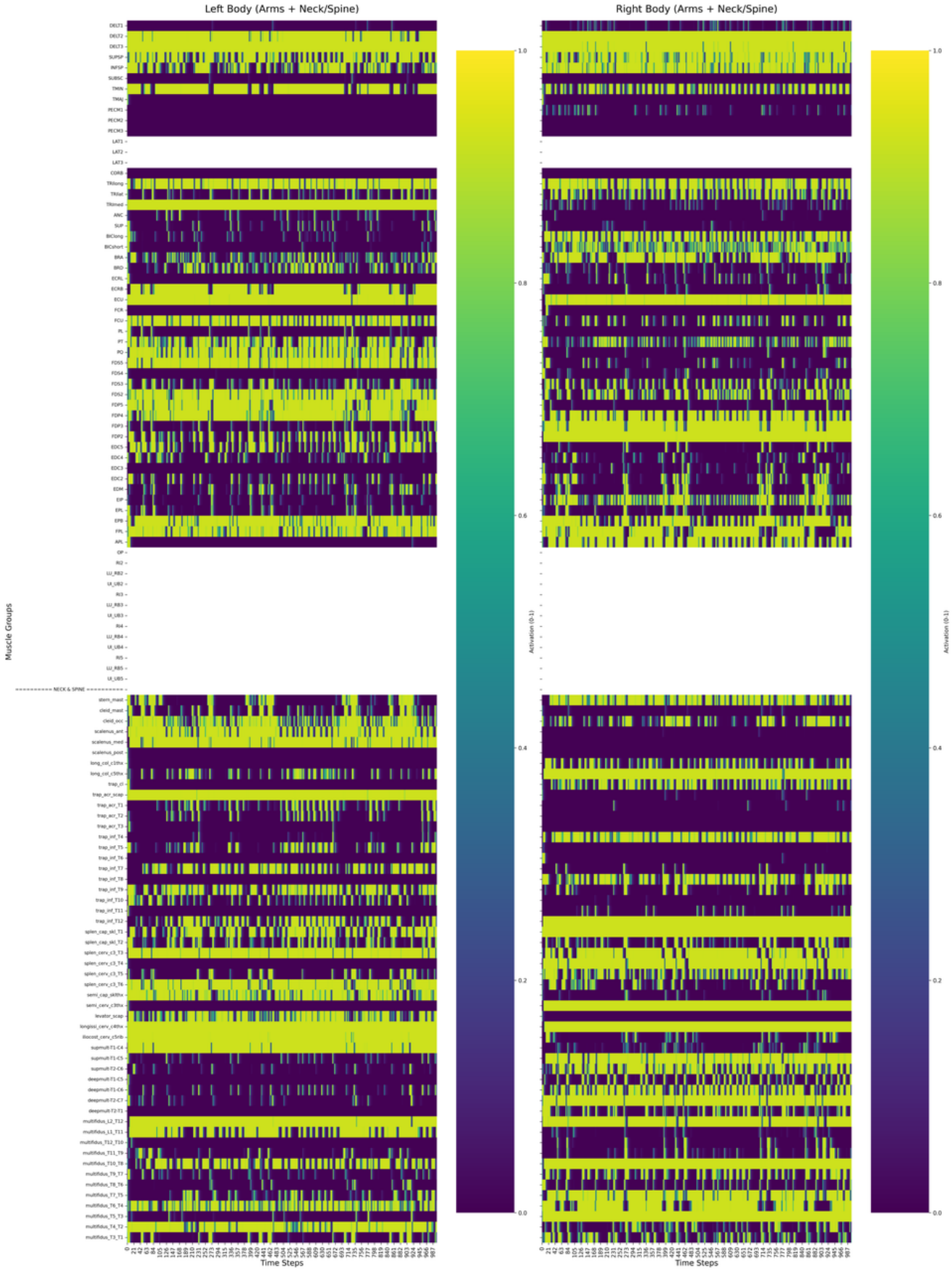

Fig. 16: **Expanded muscle activation heatmap for the 700-muscle model (Powerlift).** The top section includes the upper limb muscles, where the distinct ”blank gaps” represent the fine-grained scapular and arm muscles exclusive to the 700-model. The bottom section explicitly details the heavy recruitment of the newly available neck and spine axial stabilizers (e.g., *multifidus*, *splenius*).

2. Hard-mode Action Filtering in RL: During reinforcement learning, the policy outputs a raw action $a_t^{\mathrm{RL}} \in [-1,1]^M$. This action is first normalized to the muscle activation range:

$$a_{t,01}^{\mathrm{RL}} = \frac{a_t^{\mathrm{RL}} + 1}{2}. \tag{20}$$

The pretrained encoder and decoder then generate an expert-informed filtered action:

$$\hat{a}_t = \mathrm{Decoder}\left(\mathrm{Encoder}(a_t^{\mathrm{RL}}), P_t, S_t, M_t\right), \tag{21}$$

where $\hat{a}_t \in [0,1]^M$ represents the anatomical action prior predicted by the transformer decoder.

In hard mode, the final executable action is clipped within a fixed-radius action tube around the expert-informed action:

$$a_{t,01}^{\mathrm{env}} = \mathrm{clip}\left(a_{t,01}^{\mathrm{RL}}, \hat{a}_t - \delta, \hat{a}_t + \delta\right), \tag{22}$$

where δ is the tube radius controlling the maximum allowed deviation from the anatomical prior. The action is then mapped back to the environment action range:

$$a_t^{\mathrm{env}} = 2a_{t,01}^{\mathrm{env}} - 1. \tag{23}$$

This hard constraint prevents the policy from producing physiologically implausible muscle activations while still allowing local exploration around the expert action manifold.

3. Deviation Penalty: To further discourage unnecessary deviation from the expert-informed action, the hard mode introduces a penalty based on the violation between the original RL action and the clipped safe action:

$$\mathcal{L}_{\mathrm{dev}} = \left\| a_{t,01}^{\mathrm{RL}} - a_{t,01}^{\mathrm{env}} \right\|^2. \tag{24}$$

The environment reward is modified as:

$$R_t' = R_t - \lambda_t \mathcal{L}_{\mathrm{dev}}, \tag{25}$$

where λ_t is a decaying coefficient. At the early stage of training, the deviation penalty strongly guides the policy toward the expert action manifold. As training progresses, the penalty gradually decreases, allowing the policy to adapt more freely to task-specific objectives.

4. Transformer-based Muscle Attention: The Full Anatomical Transformer models inter-muscle dependencies through stacked Muscle Attention Blocks. Each block consists of multi-head self-attention, layer normalization, feed-forward transformations, and residual connections. By attending across muscles, the decoder captures global muscle synergy patterns and joint-level coordination, which are difficult to model using independent muscle-wise mappings.

5. Overall Principle: In hard mode, Ltent-Action RL can be viewed as reinforcement learning inside an expert-guided anatomical action tube. The pretrained encoder-decoder provides a biologically plausible action manifold, while reinforcement learning performs bounded exploration around this manifold. This mechanism reduces the difficulty of high-dimensional muscle control, suppresses unstable or energy-wasting actions, and improves the stability and physiological plausibility of learned humanoid behaviors.

H. EMG Data Processing and Comparative Analysis

This section describes the processing and comparative analysis of electromyography (EMG) signals for evaluating simulated and RL-generated muscle activations against human reference data.

1. Data Normalization and Gait Cycle Alignment EMG signals are first normalized to the gait cycle. Each signal is resampled to 101 points per cycle using cubic interpolation:

$$\tilde{x} = \mathrm{interp1d}(x_{\mathrm{old}}, x_{\mathrm{raw}}, \mathrm{kind='cubic'})(x_{\mathrm{new}}), \tag{26}$$

where x_{old} represents the original sample points and x_{new} is the normalized gait percentage from 0% to 100%. Signals are then min-max normalized to $[0,1]$:

$$x_{\mathrm{norm}} = \frac{x - x_{\mathrm{min}}}{x_{\mathrm{max}} - x_{\mathrm{min}} + \epsilon}, \tag{27}$$

with ϵ a small constant to prevent division by zero.

2. Multi-Cycle Averaging For EMG recordings with multiple gait cycles, individual cycles are extracted by detecting peaks corresponding to gait events (e.g., heel strikes). Each cycle is normalized and resampled, then averaged across cycles to produce a mean activation profile. The standard deviation across cycles is also computed and normalized by the peak-to-peak amplitude to quantify variability:

$$\sigma_{\mathrm{norm}} = \frac{\mathrm{std}(x_{\mathrm{cycles}})}{x_{\mathrm{max}} - x_{\mathrm{min}}}. \tag{28}$$

3. Simulation-Human Alignment Simulated EMG signals are aligned with the human reference by cyclically shifting the normalized simulation profile to maximize the Pearson correlation coefficient r:

$$r = \frac{\mathrm{cov}(\tilde{x}_{\mathrm{sim}}, \tilde{x}_{\mathrm{human}})}{\sigma_{\mathrm{sim}}\sigma_{\mathrm{human}}}. \tag{29}$$

The shift that maximizes r is applied to both the mean and standard deviation of the simulated signal, ensuring phase alignment with the human gait cycle.

4. Comparative Analysis For each target muscle, the correlation coefficient r quantifies temporal fidelity of the simulated or RL-generated activation relative to human EMG. The normalized standard deviation across gait cycles measures intra-cycle variability, allowing assessment of both timing and amplitude consistency. These metrics provide a quantitative basis for comparing different simulation or RL methods with human muscle activation patterns.

I. Muscle Recruitment Categorization

To analyze whole-body muscle recruitment patterns, we categorize the 416 muscle actuators into anatomically and functionally meaningful muscle families. This grouping converts the high-dimensional muscle activation vector into an interpretable recruitment profile, allowing us to quantify how different body regions contribute to each learned behavior.

For each trained policy, we perform a rollout in the corresponding musculoskeletal environment and record the activation of each muscle actuator at every simulation timestep. Let $a_{m,t}$ denote the activation of the m-th muscle at timestep t,

where $m \in \{1, \ldots, M\}$ and $M = 416$. The mean activation of each muscle over one rollout is computed as:

$$\bar{a}_m = \frac{1}{T} \sum_{t=1}^{T} a_{m,t}, \tag{30}$$

where T denotes the number of simulated timesteps before termination or the predefined maximum rollout length.

Each muscle actuator is assigned to an anatomical family according to its actuator name. Specifically, each category contains a set of predefined name keywords. A muscle is assigned to category C_k if its actuator name contains at least one keyword belonging to that category:

$$m \in C_k \quad \Longleftrightarrow \quad \exists q \in \mathcal{Q}_k \text{ s.t. } q \subseteq \text{name}(m), \tag{31}$$

where $\mathcal{Q}_k$ denotes the keyword set of the k-th muscle category.

The absolute activation mass of each muscle family is then computed by summing the mean activations of all muscles assigned to that family:

$$A_k = \sum_{m \in C_k} \bar{a}_m. \tag{32}$$

The total activation mass across all muscle families is:

$$A_{\text{total}} = \sum_{k=1}^{K} A_k. \tag{33}$$

Finally, the relative contribution of each muscle family is reported as:

$$P_k = \frac{A_k}{A_{\text{total}}} \times 100\%. \tag{34}$$

The full anatomical grouping rule used in our analysis is summarized in Table XIV. This categorization covers the major trunk, upper-limb, lower-limb, and hand muscle groups in the 416-muscle model.

Muscles that do not match any predefined keyword are assigned to an *Unclassified* group. During analysis, we monitor the activation mass of this group to verify that the predefined anatomical categories cover the majority of meaningful muscle activity. If the activation mass of the unclassified group is negligible, the resulting recruitment distribution can be considered a reliable summary of whole-body muscle usage.

This categorization allows us to compare policies not only by task reward or success rate, but also by biomechanical recruitment strategy. For example, an efficient powerlifting policy should exhibit stronger recruitment of the posterior chain and trunk stabilizers, while a locomotion policy should show coordinated contributions from hip extensors, quadriceps, hamstrings, and plantar flexors.

J. Detailed Descriptions of Residual RL Policy

To adapt the base tracking policy for robust locomotion and overcome the limitations of strict kinematic imitation across various terrains, we introduce several critical modifications to the standard Residual Reinforcement Learning (RRL) framework. These adaptations allow the highly overactuated musculoskeletal humanoid to overcome geometric discrepancies between the reference trajectory and the physical environment.

1) Action Space and Residual Scaling: Instead of allowing the residual policy to fully overwrite the base policy, we constrain the residual action space to ensure stability. The final blended action a_t sent to the musculoskeletal simulation is defined as:

$$a_t = \text{clip}\left(a_{\text{base}} + \alpha \cdot \text{clip}(a_{\text{res}}, -1.0, 1.0), -1.0, 1.0\right) \tag{35}$$

where a_t is the final blended action at time step t, a_{base} is the action inferred by the frozen base policy, a_{res} is the output of the residual network, and α is the residual scaling factor. This control ratio grants the residual policy sufficient authority to adjust muscle activations while preserving the fundamental synergistic patterns learned by the base policy. Furthermore, we deliberately disable standard kinematic penalties (e.g., action rate and activation energy penalties) to encourage aggressive exploration, retaining only a minimal L2 regularization on the residual action ($c_{\text{res}} = -0.02$).

2) Adaptations for Stair Navigation: A fundamental challenge in mimicking flat-ground or mismatched reference trajectories on physical stairs is the accumulation of vertical errors. If penalized for absolute height discrepancies, the policy prematurely terminates.

Z-Axis Agnostic Imitation Objective: We address this by masking the Z-axis displacement in the root position error calculation:

$$d_{\text{root}} = \|\mathbf{p}_{xy} - \hat{\mathbf{p}}_{xy}\|_2^2 \tag{36}$$

where d_{root} is the calculated root position error, and $\mathbf{p}_{xy}$ and $\hat{\mathbf{p}}_{xy}$ are the 2D planar coordinates of the simulated and reference root positions, respectively. This height-invariant formulation allows the humanoid to step onto elevated platforms without suffering catastrophic drops in the tracking reward.

Elastic Trajectory Pacing: To prevent the policy from being dragged forward when encountering physical obstacles (e.g., tripping on a stair edge), we decouple the reference step from the physical simulation time. The environment implements an *elastic tracking* mechanism: the reference frame index advances only if the humanoid's root deviation remains within a safe threshold ($d < 0.4$ m). If the agent lags, the trajectory progression slows down by 50%, explicitly waiting for the agent to recover its balance.

Task-Driven Progress Reward: To break the deadlock of purely conservative imitation, we augment the imitation reward with a task-driven progress reward:

$$\begin{aligned} r_{\text{prog}} = w_{\text{prog}} \Big(& \text{clip}(v_{\text{proj}}, 0, 2.0) \\ & + 0.5 \cdot \text{clip}(v_z, 0, 1.5) \Big) \end{aligned} \tag{37}$$

where r_{prog} is the calculated progress reward, $w_{\text{prog}} = 0.3$ is the weight coefficient for this reward, v_{proj} is the current root velocity projected onto the directional vector pointing to the future reference pelvis position (1.0s ahead), and v_z is the absolute vertical climbing velocity.

TABLE XIV: Anatomical muscle-family categories used for whole-body recruitment analysis. Each actuator is assigned to a category according to whether its name contains one of the corresponding keywords.

Body Region	Muscle Category	Actuator Name Keywords
Core and Back	Paraspinals (Deep)	`Ps_`
	Iliocostalis	`IL_`
	Longissimus (Erector Spinae)	`LTpT`, `LTpL`
	Abdominals	`rect_abd`, `EO`, `IO`
	Other Trunk/Back	`LAT`, `LD`, `QL`, `MF_m`, `TRAP`, `RHOM`
	Pectorals	`PEC`
Upper Extremity	Deltoids	`DELT`
	Rotator Cuff	`SUPSP`, `INFSP`, `SUBSC`, `TMIN`, `TMAJ`, `CORB`
	Elbow Flexors	`BIC`, `BRA`, `BRD`
	Elbow Extensors	`TRI`, `ANC`
	Forearm Pronators/Supinators	`SUP`, `PT`, `PQ`
	Wrist/Finger Flexors	`FCR`, `FCU`, `PL`, `FDS`, `FDP`, `FPL`
	Wrist/Finger Extensors	`ECR`, `ECU`, `EDC`, `EDM`, `EIP`, `EPL`, `EPB`, `APL`
	Hand Intrinsics	`OP`, `RI2`, `RI3`, `RI4`, `RI5`, `LU_`, `UI_`
Lower Extremity	Glutes (Hip Extensors)	`GMAX`, `GMED`, `GMIN`, `PIR`, `GEM`, `OINT`, `QFM`, `glmax`, `glmed`, `glmin`, `piri`
	Quadriceps (Knee Extensors)	`VAS`, `RF`, `VI`, `VL`, `VM`, `vasint`, `vaslat`, `vasmed`, `recfem`
	Hip Flexors/Adductors	`ILIACUS`, `PSOAS`, `ADD`, `PECT`, `GRAC`, `TFL`, `iliacus`, `psoas`, `add`, `grac`, `tfl`, `sart`
	Hamstrings (Knee Flexors)	`BFLH`, `BFSH`, `SEMIM`, `SEMIT`, `bflh`, `bfsh`, `semiten`, `semimem`
	Calves (Plantar Flexors)	`GAS`, `SOL`, `soleus`, `gaslat`, `gasmed`
	Other Leg/Foot	`TA`, `EDL`, `EHL`, `FDL`, `FHL`, `TP`, `PER`, `edl`, `ehl`, `fdl`, `fhl`, `perbrev`, `tibant`, `tibpost`, `perlong`

3) Enhancements for Energy-Efficient Walking: While navigating stairs requires aggressive spatial exploration, flat-ground walking prioritizes energy efficiency and stable transitions between stance and swing phases. Strict kinematic imitation often forces the musculoskeletal model into stiff, metabolically inefficient gaits.

Phase-Dependent Residual Authority: We dynamically modulate the residual network's authority based on the gait phase. The swing leg requires more flexibility to adjust foot clearance, whereas the stance leg demands strict synergistic stability. The control ratio is formulated as:

$$\alpha_i(t) = \alpha_{\text{stance}} + (\alpha_{\text{swing}} - \alpha_{\text{stance}}) \cdot \mathcal{I}_{\text{swing}}(i, t) \tag{38}$$

where $\alpha_i(t)$ represents the dynamic residual scaling factor for limb i at time step t, i denotes the respective limb, $\alpha_{\text{stance}} = 0.2$ and $\alpha_{\text{swing}} = 0.6$ are the base control ratios for the stance and swing phases respectively (the latter is used for the stair task as well), and $\mathcal{I}_{\text{swing}}(i, t)$ is a boolean indicator function derived from foot-ground contact sensors.

Metabolic-Aware Muscle Regularization: To eliminate high-frequency, non-functional muscle twitches, we reinstate a smoothed metabolic penalty specifically for the walk task:

$$p_{\text{met}} = -w_{\text{met}} \sum_{j=1}^{N} \left(a_{j,t}^2 + \lambda (a_{j,t} - a_{j,t-1})^2 \right) \tag{39}$$

where p_{met} is the metabolic penalty, $N = 416$ is the total number of muscles, $a_{j,t}$ is the activation of muscle j at the current time step t, $a_{j,t-1}$ is the activation of the same muscle at the previous time step, $w_{\text{met}} = 0.05$ is the scaling weight for the penalty, and $\lambda = 0.1$ is the smoothing coefficient. This ensures the learned walking gait is kinematically accurate and biologically plausible.

4) Adaptations for High-Impact Locomotion (Run Task): Running fundamentally differs from walking due to the presence of an aerial flight phase. Discrepancies in ground contact timing between the reference kinematic data and the physics simulation often cause failures during impact.

Flight-Phase Propulsive Reward: During the aerial phase, we temporarily relax the tracking penalty by 50% and introduce a propulsive velocity reward:

$$r_{\text{propel}} = w_{\text{vel}} \cdot \max(0, v_x - \hat{v}_x) \cdot \mathcal{I}_{\text{air}} \tag{40}$$

where r_{propel} is the propulsive velocity reward, w_{vel} is the weight coefficient for this reward, v_x is the current forward velocity, $\hat{v}_x$ is the target running speed, and $\mathcal{I}_{\text{air}}$ indicates the flight phase. This encourages explosive muscle forces during toe-off.

Ground Reaction Force (GRF) Smoothing: High-speed impacts cause severe numerical instability in the simulation. We introduce a soft penalty for excessive vertical Ground Reaction Forces (GRF):

$$p_{\text{impact}} = -w_{\text{grf}} \cdot \max(0, F_z - F_{\text{limit}}) \tag{41}$$

where p_{impact} is the soft impact penalty, w_{grf} is the weight coefficient for the GRF penalty, F_z is the total measured vertical contact force, and F_{limit} is set to $2.5\times$ the humanoid's

body weight. This forces the residual policy to learn active shock absorption strategies prior to touchdown.

K. Extended Discussion on Limitations and Future Directions

While the Musculoskeletal Humanoid Benchmark provides a comprehensive and diverse foundation for muscle-driven control, it is essential to acknowledge several limitations in the current scope of this work. These limitations also highlight promising avenues for future research in biomechanical reinforcement learning and humanoid robotics.

1) Absence of Dexterous Manipulation: The current iteration of our benchmark primarily concentrates on gross motor skills, such as full-body locomotion, postural stabilization, and basic environmental interactions (e.g., pushing a door or reaching). However, one of the most defining characteristics of the human musculoskeletal system is its extraordinary capacity for fine motor control, particularly in the hands and fingers. Modeling the biomechanics of the human hand is highly complex due to the dense routing of tendons, high degrees of freedom, and intricate muscle synergies. Consequently, our benchmark currently lacks fine-grained dexterous manipulation tasks, such as in-hand object manipulation, tool use, or precise grasping. Future work should integrate high-fidelity musculoskeletal hand models into the full-body framework, enabling the study of how gross arm movements dynamically couple with fine finger manipulations.

2) Transitioning to Compositional and Long-Horizon Tasks: Currently, our benchmark evaluates policies on isolated, single-objective tasks (e.g., exclusively walking, or solely executing a powerlift). Real-world humanoid operation, however, inherently demands the execution of compositional and long-horizon behaviors. For instance, a realistic scenario might require the robot to walk toward a chair, pick up a barbell along the way, and subsequently perform a seated lift. Addressing such long-horizon tasks will require policies to not only master atomic skills but also smoothly transition between them without losing structural balance or experiencing catastrophic energy depletion. Future iterations of this benchmark will aim to include compositional task suites, serving as a testbed for hierarchical reinforcement learning (HRL), skill discovery, and long-term fatigue modeling in muscle-driven systems.

3) Constraints in Comprehensive Physiological Validation: A core motivation of musculoskeletal modeling is to generate biologically plausible behaviors. However, our rigorous physiological validation—specifically the comparison between policy-generated muscle activations and ground-truth human Electromyography (EMG) signals—is currently restricted predominantly to walking tasks. This limitation stems from a scarcity of comprehensive, open-source EMG datasets for highly dynamic or unconventional tasks (e.g., hurdle clearance, whole-body crawling, or wobble board balancing). To achieve a more universal physiological benchmark, future research must involve establishing advanced data collection pipelines—synchronizing high-density EMG with full-body motion capture—to build a diverse dataset covering a broader spectrum of human motor skills.

4) Sim-to-Real Gap and Hardware Validation: Finally, all evaluations within this benchmark are conducted in a simulated physics environment. Although the simulation enforces rigorous biomechanical and physical constraints, it inevitably abstracts away certain real-world complexities, such as sensor noise, communication latency, and unmodeled actuator dynamics. Validating these muscle-driven policies on physical hardware remains a formidable challenge. Real-world musculoskeletal humanoids (e.g., tendon-driven robots) are notoriously difficult to build and maintain. Future directions include bridging the sim-to-real gap by either deploying these policies on state-of-the-art tendon-driven hardware or exploring methodologies to map virtual muscle activations (acting as variable stiffness/compliance signals) onto traditional electric-motor-driven humanoids equipped with whole-body impedance control.