Proceedings of Machine Learning Research vol vvv:1–19, 2026

Embodied Learning of Reward for Musculoskeletal Control with Vision Language Models

Saraswati Soedarmadji[1] CHENXUYING24@MAILS.TSINGHUA.EDU.CN
Yunyue Wei[1] YUNYUEWEI@MAIL.TSINGHUA.EDU.CN
Chen Zhang[1] CZHANG.EMAIL@GMAIL.COM
Yisong Yue[2] YYUE@CALTECH.EDU
Yanan Sui[1] YSUI@TSINGHUA.EDU.CN
[1]*Tsinghua University,* [2]*California Institute of Technology*

Abstract

Designing effective reward functions is a fundamental challenge for controlling high-dimensional human musculoskeletal systems. For example, humans can describe movement goals like "walk forward with upright posture", but the underlying motor strategies that realize these goals are implicit and complex. We introduce Motion from Vision-Language Representation (*MoVLR*), which uses vision-language models (VLMs) to bridge natural language descriptions and human motion in musculoskeletal control. Instead of handcrafted rewards, *MoVLR* integrates control learning with VLM feedback to align control policies with physically coherent behaviors. Our approach transforms language and visual assessments into guidance for embodied learning of a variety of human movements from high-level descriptions. *MoVLR* automatically designs and optimizes rewards for the control of a high-dimensional musculoskeletal model for manipulation and locomotion. These results indicate that vision-language models can effectively ground abstract motion descriptions in the implicit principles of physiological motor control.

Keywords: Embodied Learning, Internal Dynamics, Musculoskeletal Control, Motion Representation, Vision Language Models

1. Introduction

Humans acquire motion control through practice, imitation, and external guidance, with effective control arising from the intricate interactions between the nervous and musculoskeletal systems. Unlike general robotic systems, musculoskeletal agents exhibit highly nonlinear, overactuated, and high-dimensional dynamics (Tohidi et al., 2016). Multiple muscles or synergies can produce identical joint motions, while coordinated body movements demand precise and elaborate control of the entire musculature. These factors present fundamental challenges to achieving efficient natural motion control. Recent progress in high-dimensional musculoskeletal control has utilized learning-based frameworks (Schumacher et al., 2022; He et al., 2024). However, most existing methods rely on heuristic learning objectives such as velocity tracking or energy minimization, failing to capture the nuanced structure of motion complexity. Although such rewards can enable basic task completion, they inadequately embody anatomical principles and often yield biomechanically unnatural movements.

Recent advances in large language models (LLMs) and vision-language models (VLMs) have shown promise in eliciting high-level notions of movement quality and coordination to design executable reward designs, enabling automated reward learning (Ma et al., 2023; Zeng et al., 2024).

Despite encouraging results across several robotic learning tasks, most works rely on episodic statistics as static feedback for reward learning. Their evaluations are limited in low-dimensional, torque-driven systems with explicit dynamics. Meanwhile, language can convey explicit instructions for human motion, yet motor learning also largely depends on implicit sensorimotor patterns that are difficult to formalize. Therefore, whether LLMs/VLMs can be leveraged to infer internal dynamical representations for motion reward learning of high-dimensional musculoskeletal control remains an open question.

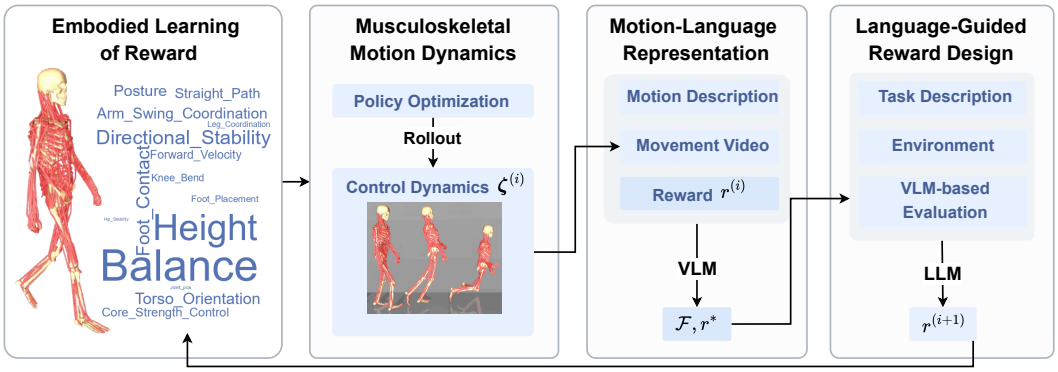


Figure 1: Workflow of *MoVLR*. Policy optimization is performed to provide high-dimensional musculoskeletal dynamics of the reward candidate. A VLM evaluates the corresponding movement video $\zeta^{(i)}$ to update the current best reward design r^* and suggest biomechanical improvements $\mathcal{F}$ for a LLM to refine reward generation of $r^{(i+1)}$.

In this paper, we present Motion from Vision Language Representation (*MoVLR*), which automatically learns rewards for high-dimensional musculoskeletal control by integrating both descriptive and dynamical feedback. As shown in Figure 1, *MoVLR* extracts high-dimensional musculoskeletal dynamics of candidate rewardsl via policy optimization and rollout. The resulting control dynamics are rendered into movement videos to provide dynamical feedback. A vision–language model then evaluates the agent's motion and produces structured biomechanical feedback, which is used by a LLM to refine the reward generation process. By encoding temporal dynamics into semantically meaningful descriptors, *MoVLR* bridges raw perceptual input with domain-informed motion representations, providing a scalable path toward biomechanically realistic control reward for high-dimensional musculoskeletal systems. The supplementary material and complete code to reproduce all experimental results can be found at: https://sites.google.com/view/movlr/home.

Contributions: We develop *MoVLR*, a fully automatic pipeline which effectively capture implicit dynamics and design explicit rewards for controlling high-dimensional musculoskeletal systems. We show that *MoVLR* generalizes across movement types, environments and system morphologies with explainable reward terms to represent implicit musculoskeletal dynamics. Our work provides a novel approach that enables interpretable evaluation of motor performance, adaptive refinement of reward design through vision-language feedback, and a transferable control framework for natural and coordinated motion.

2. Related Works

2.1. Control of musculoskeletal systems

The control of high-dimensional musculoskeletal systems presents a fundamental challenge due to the high dimensionality and redundancy inherent in human-like actuation. Considerable progress has been made in building faithful simulation environments for muscle-tendon dynamics and joint kinematics, enabling more realistic learning and evaluation (Lee et al., 2019; Song et al., 2021; Caggiano et al., 2022). To manage the control complexity, many works simplified training by decomposing the process into hierarchical pipelines (Lee et al., 2019; Park et al., 2022; Feng et al., 2023) or by employing curriculum learning schedules (Caggiano et al., 2023; Park et al., 2025). Other studies tried to improve sample efficiency via bio-inspired sampling (Schumacher et al., 2022), coordinated latent-space exploration (Chiappa et al., 2023; Simos et al., 2025), or model-based control (Hansen et al., 2023). Recent approaches further reduced control dimensionality by extracting muscle synergies informed by task demands or anatomy (Berg et al., 2024; He et al., 2024). Despite these advances, the resulting performance often hinges on manually engineered reward functions that are crafted with human effort to accomplish specific tasks and encode implicit domain knowledge.

2.2. Language and multimodal driven reward design

Recent advances in large language models have demonstrated their ability to facilitate reward and feedback design for robotics and simulation systems (Goyal et al., 2019; Song et al., 2023; Ma et al., 2024; Xie et al., 2024; Masadome and Harada, 2025). Eureka (Ma et al., 2023) uses code-generating LLMs to synthesize dense, human-level reward functions that surpass manually engineered counterparts in both expressivity and task relevance. While originally proposed in the context of reinforcement learning, this paradigm is equally relevant to control systems: the generated signals can be viewed as structured feedback that shapes system dynamics toward desired trajectories (Guo et al., 2024; Narimani and Emami, 2025; Zhou et al., 2025).

Beyond text-only models, recent VLM-based works have further demonstrated the value of incorporating multimodal signals such as images or video into feedback design (Rocamonde et al., 2023; Brohan et al., 2023; Ge et al., 2023; Wang et al., 2024; Zeng et al., 2024), suggesting a trend towards more expressive and physically grounded control objectives. For example, HARMON (Jiang et al., 2024) leverages a VLM to iteratively refine humanoid motion by evaluating sequences of rendered frames against language descriptions, using these visual snapshots to assess semantic alignment and guide motion adjustments. However, existing methods still lack a principled way of turning VLM/LLM feedback into structured dynamical signals that can shape the reward function, and often provide only coarse or ad hoc representations of multimodal feedback rather than integrating it systematically into control.

3. Preliminaries

3.1. High-dimensional musculoskeletal control

Musculoskeletal systems. In this paper, our target systems are high-dimensional, over-actuated musculoskeletal systems with dynamics governed by

$$\boldsymbol{M}(\boldsymbol{q})\,\ddot{\boldsymbol{q}} + \boldsymbol{c}(\boldsymbol{q},\dot{\boldsymbol{q}}) = \boldsymbol{J}_m^\top \boldsymbol{f}_m + \boldsymbol{J}_c^\top \boldsymbol{f}_c + \boldsymbol{\tau}_{\text{ext}}, \tag{1}$$

where $\boldsymbol{q}$ are generalized joint coordinates, $\boldsymbol{M}(\boldsymbol{q})$ is the mass matrix, and $\boldsymbol{c}(\boldsymbol{q}, \dot{\boldsymbol{q}})$ represents Coriolis and gravitational effects. The Jacobians $\boldsymbol{J}_m$ and $\boldsymbol{J}_c$ map actuator and constraint forces ($\boldsymbol{f}_m, \boldsymbol{f}_c$) to generalized coordinates, and τ_{ext} denotes external torques from the environment. Muscles are modeled as first-order actuators driven by neural controls $\boldsymbol{u}$ with activation $\boldsymbol{a}$, where the force f_m generated by one actuator is formulated by:

$$f_m = F_k(l, v)\, a + F_p(l), \quad \frac{\partial a}{\partial t} = \frac{u - a}{\tau(u, a)}, \tag{2}$$

with actuator length l, velocity v,gains F_k, bias F_p and time coefficient τ. Note that F_k, F_p and τ vary with muscle states, leading to high non-linearity. In our experiments, we use MS-Human-700 model (Zuo et al., 2024) as the major benchmark for human full-body musculoskeletal control. The model consists of 206 joints and 700 muscle-tendon actuators. Additional experiments can involve other morphologies.

Policy optimization problem. We model the high-dimensional musculoskeletal control problem as a finite horizon Markov decision process (MDP) with state $\boldsymbol{s} \in \mathcal{S}$, control $\boldsymbol{u} \in \mathcal{U}$, dynamics $\boldsymbol{f} \coloneqq \mathcal{S} \times \mathcal{U} \to \mathcal{S}$, horizon T, policy $\boldsymbol{\pi} \coloneqq \mathcal{S} \to \mathcal{U}$ and reward $r \coloneqq \mathcal{S} \times \mathcal{U} \to \mathbb{R}$. In this paper, we formulize the reward as linear combination of reward terms: $r = \boldsymbol{w} \cdot \boldsymbol{r}$ with $\boldsymbol{w} = (w_1, w_2, \cdots)$ and $\boldsymbol{r} = (r_1, r_2, \cdots)$ as the weights and values of specific reward terms. For example, a reward for human walking can be expressed as:

$$r_{\text{walk}} = w_{\text{height}} r_{\text{height}} + w_{\text{balance}} r_{\text{balance}} + \cdots + w_{\text{forward}} r_{\text{forward}}. \tag{3}$$

Given initial state $\boldsymbol{s}_0$, we aim to achieve stable control of the system by finding a policy $\boldsymbol{\pi}^*$ that maximize the reward function:

$$\boldsymbol{\pi}^* = \text{argmax}_{\boldsymbol{\pi}} \sum_{t=0}^{T-1} r(\boldsymbol{s}_t, \boldsymbol{u}_t), \quad \boldsymbol{u}_t = \boldsymbol{\pi}(\boldsymbol{s}), \quad \boldsymbol{s}_{t+1} = \boldsymbol{f}(\boldsymbol{s}_t, \boldsymbol{u}_t). \tag{4}$$

The above policy optimization problem is also equivalent to reward minimization commonly used in control-based methods, where the reward function is the negative reward.

3.2. Reward learning for musculoskeletal control

While the above control problem provides single-step reward definition, the control performances are usually evaluated over full horizon. The objective of reward learning is to find single step reward r^* that maximize the global reward function R:

$$r^* = \text{argmax}_r R(\boldsymbol{\zeta}), \quad \boldsymbol{\zeta} = (\boldsymbol{s}_0, \boldsymbol{a}_0, \cdots, \boldsymbol{s}_{T-1}, \boldsymbol{a}_{T-1}), \quad \boldsymbol{u}_t = \boldsymbol{\pi}_r^*(\boldsymbol{s}_t), \quad \boldsymbol{s}_{t+1} = \boldsymbol{f}(\boldsymbol{s}_t, \boldsymbol{u}_t), \tag{5}$$

where $\boldsymbol{\pi}_r^*$ is the optimal policy derived by maximizing r. In practice, the global reward R can be high-level as motion descriptions in natural language, such as "walk forward" or "grab the bottle". The single-step reward r consists of multiple reward terms and parameters as code pieces which need to be compatible with the policy optimization framework. Achieving effective reward learning requires: (1) **Embodied understanding of human movement**: extracting implicit biomechanical knowledge of the musculoskeletal system based on the motion description, and (2) **Effective reward generation:** integrating multimodal feedback to design explainable reward terms which are executable for policy optimization.

4. Methods

To address the challenges of limited task understanding and multimodal reasoning, we propose *MoVLR*, a control-in-the-loop framework that integrates large language models (LLMs) and vision–language models (VLMs) into the reward learning process. *MoVLR* bridges **explicit behavior specifications** expressed in natural language with the **implicit dynamical representations** essential for effective control. The key idea is to incorporate video observations of policy-executed trajectories into the iterated learning loop, enabling the model to jointly reason over linguistic intent and physical motion. This multimodal feedback provides structured insights into trajectory feasibility, biomechanical coherence, and task completion, ultimately yielding reward functions that are more aligned with the underlying system dynamics.

Algorithm 1 *MoVLR*

Require: Motion description R, environment code $\mathcal{E}$, max iterations N, initial reward design $r^{(0)}$

1: $\boldsymbol{\zeta}^* \leftarrow \emptyset$, $r^* \leftarrow \emptyset$
2: **for** $i = 0, \cdots, N-1$ **do**
▷ `Musculoskeletal motion dynamics`
3: Obtain $\boldsymbol{\pi}^*_{r^{(i)}}$ by optimizing e.q. (4)
4: Obtain $\boldsymbol{\zeta}^{(i)}$ by rollout $\boldsymbol{\pi}^*_{r^{(i)}}$
▷ `Motion-language representation`
5: $\boldsymbol{\zeta}^*, r^* \leftarrow \text{VLM}(R, \boldsymbol{\zeta}^{(i)}, \boldsymbol{\zeta}^*, r^{(i)}, r^*)$
6: $\mathcal{F} \leftarrow \text{VLM}(R, \boldsymbol{\zeta}^*, r^*)$
▷ `Language-guided reward design`
7: $r^{(i+1)} \sim \text{LLM}(R, \mathcal{E}, \mathcal{F}, r^*)$
8: **end for**
9: **Return:** Optimized reward r^*

The workflow of *MoVLR* is summarized in Algorithm 1. At each iteration, the policy is optimized based on the current reward proposal, producing dynamical feedback that reflects the control performance (**musculoskeletal motion dynamics**, line 3-4). Given the executed control dynamics (rendered as video), a vision–language model (VLM) performs reflective evaluation, updating the current best reward design and generating a textual summary of the observed task performance (**motion-language representation**, line 5-6). incorporates both the motion description and the VLM-generated summary from the previous iteration to refine the reward generation process (**language-guided reward design**, line 7). Below we discuss the implementation details of each components for effective reward learning of musculoskeletal control.

4.1. Musculoskeletal motion dynamics

We evaluate each reward proposal by performing policy optimization to generate dynamical control trajectories as feedback. In *MoVLR*, we adopt MPC^2 as the control policy—a model-based planner that employs a hierarchical control pipeline for musculoskeletal systems (Wei et al., 2025). Compared with reinforcement-learning based control with hours to days of training, MPC^2 employ a training-free pipeline to significantly reduces the policy optimization time to minutes, allowing more reward learning iterations in *MoVLR* (see Appendix A.1 for method details). The resulting musculoskeletal motion dynamics obtained by rolling out the optimized policy serve as the dynamical feedback for refining the reward design.

4.2. Motion-language representation

We employ the VLM as a semantic observer that produces interpretable, language-based evaluations rather than scalar scores. The VLM compares the rendered control dynamics $\boldsymbol{\zeta}^{(i)}$ against the dynamics generated under the current best reward design. If the newly proposed reward yields control sequences that better align with the motion description, both the current best reward r^* and

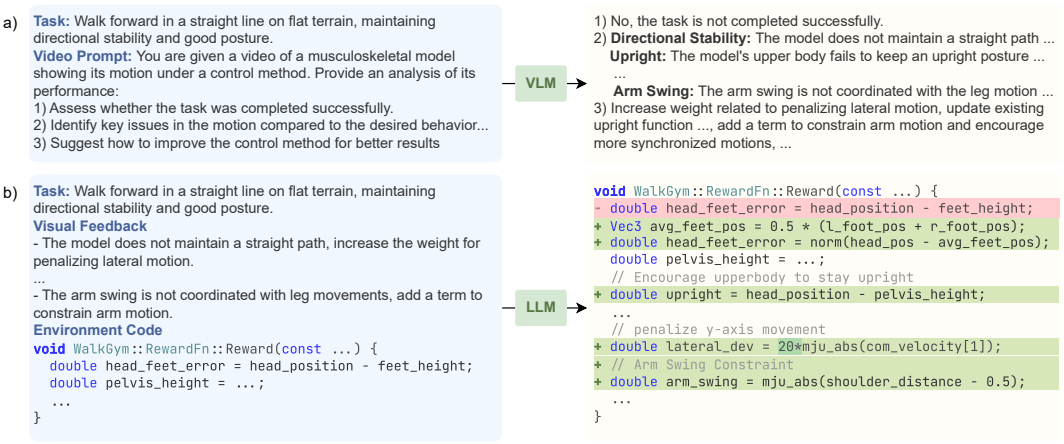


Figure 2: Example inputs and outputs of the (a) VLM, and (b) LLM. The VLM analyzes a video motion sequence based on the given motion description and provides diagnostic feedback. The LLM uses this feedback to design corresponding code modifications to the reward function.

the corresponding control dynamics ζ^* are updated. The VLM also produces structured textual feedback $\mathcal{F}$ of r^* which qualitatively evaluates the motion relative to R. This feedback serves as reflective input to the LLM, guiding subsequent iterations of reward design. Through this mechanism, *MoVLR* establishes a multimodal interface for specifying and interpreting complex motor behaviors, integrating visual and textual modalities to reason about the correspondence between natural-language motion descriptions and observed motion.

4.3. Language-guided reward design

Designing executable reward functions from multimodal inputs requires mapping semantic and structural priors to physically meaningful quantities. In musculoskeletal control, this must capture nonlinear couplings between balance, posture, and coordination—relations difficult to encode manually. We employ a language-guided reward synthesis process, where an LLM generates interpretable reward terms by reasoning over both linguistic and structural context:

(1) Motion Description. The natural-language specification R defines the high-level control objective (e.g., "make the arm grasp and lift the bottle"). It serves as a semantic prior that highlights key performance factors such as grasp stability, coordination, or smoothness.

(2) Environment. The environment $\mathcal{E}$ specifies the state, control, and transition dynamics. Parsing $\mathcal{E}$ allows the identification of physically relevant variables (e.g., joint angles, actuator lengths, and contact forces) that can parameterize executable reward terms.

Given above contextual information along with and VLM-based evaluations $\mathcal{F}$, the LLM performs a local search over the current best reward design r^* to synthesize new reward terms. Each term encodes a biomechanical sub-objective such as orientation tracking, smoothness, or joint stability. At each iteration, reward proposals are continuously designed and refined until an executable design is identified, yielding an interpretable reward function suitable for policy optimization and control-based dynamical feedback design. Compared with traditional language-based reward design approaches that rely solely on LLMs, *MoVLR* introduces dynamical control feedback, enabling

the VLM to reflect on kinematic and postural precision—an essential capability for achieving stable musculoskeletal control.

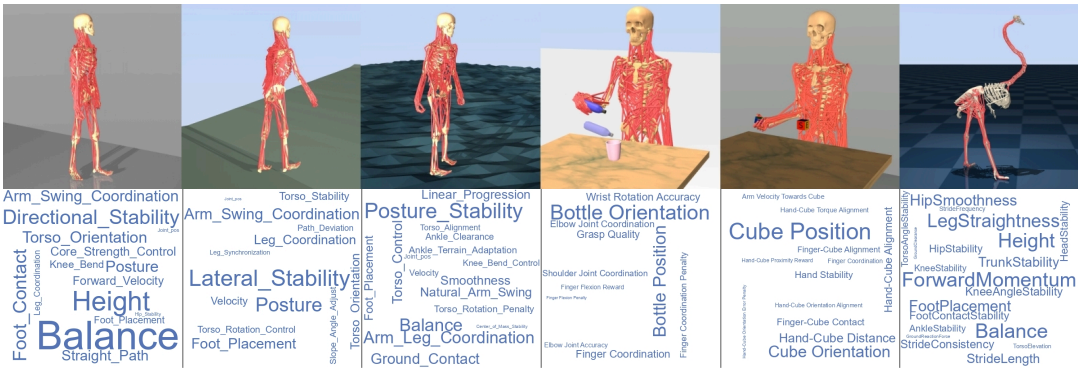


Figure 3: Overview of the six evaluated tasks. The top row illustrates the environment setup for each task, and the bottom row visualizes the relative weighting of learned reward terms.

5. Experiments

We evaluate *MoVLR* across a diverse set of musculoskeletal systems and tasks, assessing its capacity to design reward functions, solve novel tasks, and integrate different forms of human input. Unless otherwise noted, all VLM and LLM-based reward design and feedback algorithms are built on the Gemini (Team et al., 2023) and Qwen models (Yang et al., 2024), specifically `gemini-2.0-flash` and `Qwen2.5-Coder-32B-Instruct` models.

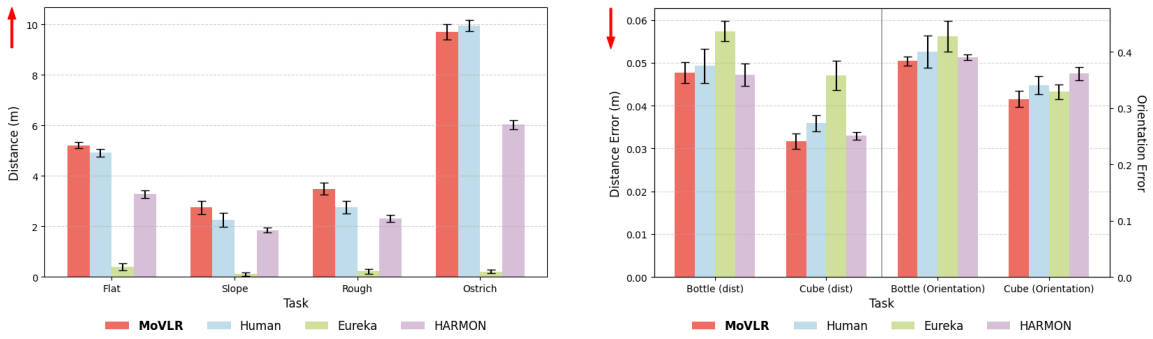
5.1. Environments

As shown in Figure 3, our experimental setup spans three musculoskeletal systems and a total of eight tasks implemented in the MuJoCo simulator (Todorov et al., 2012). The suite is designed to capture a broad spectrum of control challenges. It includes three locomotion environments (flat, rough, and sloped terrain) that test stability and adaptability under varying ground conditions. The flat terrain setting additionally includes two variants – a turning task (left/right directional transitions) and an injured-body condition where selected leg muscle groups are weakened – to evaluate gait robustness and compensatory control strategies. The suite further includes two manipulation tasks (bottle pouring and cube manipulation) that emphasize coordination and precision, and one non-human locomotion task based on an ostrich muscle model that evaluates generalization beyond human morphology.

5.2. Experimental Results

Comparison with state-of-the-art LLM/VLM based methods. We evaluate our method against three baselines: human-engineered reward functions (Wei et al., 2025), Eureka (Ma et al., 2023), and HARMON (Jiang et al., 2024) on both locomotion and manipulation tasks. Our implementation details of method and baselines are elaborated in Appendix A.2. Evaluation focuses on final per-

formance after convergence, measured by average walking distance over 10 seconds for locomotion and by object position and orientation errors for manipulation.



(a) Locomotion Tasks

(b) Manipulation Tasks

Figure 4: Performance comparison of *MoVLR* against baselines across (a) locomotion and (b) manipulation tasks. Locomotion is measured by total distance walked in 10s (higher is better), while manipulation is evaluated by object distance and orientation errors (lower is better).

As shown in Figure 4 (a), across all locomotion environments, *MoVLR* consistently produces higher task performance, yielding the longest walking distances on flat, sloped, and rough terrains. The improvements are most pronounced in challenging settings, where terrain irregularities require adaptive stability and coordinated motion. Despite slight a slightly lower performance compared to the human baseline, *MoVLR* also generalizes effectively to the ostrich environment, maintaining strong performance despite substantial morphological differences from the human model. These results demonstrate that multimodal reward refinement leads to more robust and transferable control objectives across biomechanical structures.

As shown in Figure 4 (b), *MoVLR* achieves the lowest average position and orientation errors compared to all baselines in manipulation tasks, indicating more precise and stable object interactions. The improvements are consistent across both bottle-pouring and cube-rotation movements, suggesting that multimodal feedback enhances the alignment between high-level motion intent and low-level control behavior.

Additional heatmaps for the remaining four tasks are included in appendix A.4, illustrating consistent refinement dynamics across both locomotion and manipulation domains.

Evolution of weighted reward terms across refinement stages. The progression of residual reward weights across refinement stages reveals how the feedback-driven process reorganizes the internal optimization landscape toward biomechanically consistent behavior. Visual inspection of the heatmaps shows clear temporal structure in how specific reward terms are emphasized, attenuated, or replaced as refinement proceeds. Rather than uniform or random variation, the weights evolve in a task-specific and interpretable manner that reflects the gradual integration of control priorities derived from feedback. This progression is visually illustrated in Figure 5, showing the musculoskeletal agent's transition from instability to coordinated walking as successive stages refine control priorities such as balance, posture, and stride formation.

In Figure 6 (a), we demonstrate the learning evolution of *MoVLR* in locomotion task over rough terrain. We observe early refinement stages concentrate weight on coarse global stability terms such

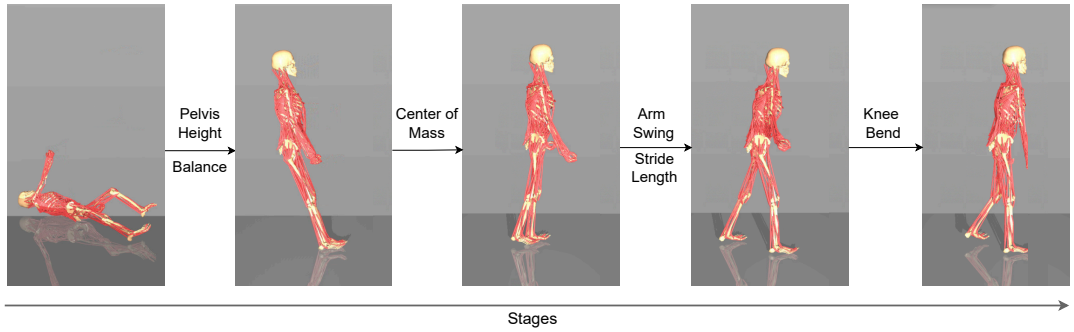


Figure 5: Progressive improvement of the musculoskeletal model's gait across training stages based on movement video.

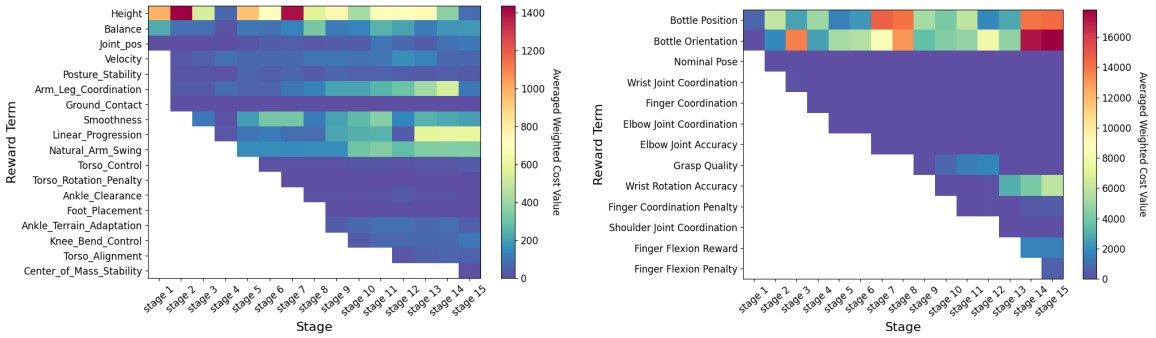
as *height*, *velocity*, and *balance*. These initial weightings dominate the first few iterations, suggesting that the system prioritizes feasibility and upright posture before attempting finer coordination. As refinement progresses, the influence of these global terms decreases steadily, while localized biomechanical descriptors, such as *foot placement*, *hip alignment*, and *knee control*, become more prominent. This redistribution indicates a shift from whole-body stabilization to detailed gait regulation. By later stages, the weight profiles become highly structured, with consistent activation around terms governing *step symmetry*, *torso orientation*, and *ankle stability*, suggesting convergence toward coordinated, rhythmic locomotion.

In the bottle pouring task shown in Figure 6 (b), a similar hierarchical refinement pattern is observed. Initial stages emphasize gross spatial alignment through terms such as *bottle position* and *bottle orientation*, enabling task feasibility. With continued refinement, weights shift toward fine motor control components, including *grasp quality, elbow joint accuracy*, and *finger coordination*. The redistribution of weights toward distal control terms indicates that the framework captures the need for precise joint coordination in achieving smooth and stable object manipulation.

Ablation Studies. To better understand the contribution of each design component and the generality of the proposed framework, we conduct a series of ablation studies examining (1) reward generalizability across environments, model conditions and policy parameterization; (2) the use of a single unified vision–language model for feedback and code generation, and

(1) Reward generalizability. We test transferring reward functions designed on flat terrain to new environments without additional refinement. The transferred rewards show strong generalization across terrains and morphologies. While performance drops moderately on rough (2.41 m vs. 2.76 m) and sloped (1.99 m vs. 2.25 m) terrains, agents remain stable and capable of sustained locomotion. In the injured-body setting, transfer performance slightly improves (5.12 m vs. 4.8 m), indicating robustness to actuator failure. The method also enables a left-turn behavior previously infeasible with hand-engineered rewards, showing that the learned reward structure extends beyond the original environment configuration.

We test reward functions learned by *MoVLR* in a reinforcement learning setting using DynSyn (He et al., 2024) for the bottle-pouring task. The transferred rewards enable successful task completion without further tuning, producing stable pouring trajectories, demonstrating that *MoVLR*-designed rewards capture generalizable structure transferable across control algorithms.



(a) Rough Terrain (b) Bottle Pouring

Figure 6: Weighted reward terms per stage for (a) locomotion task, (b) manipulation task

(2) VLM-only reward learning. The framework's design separates the vision–language feedback and the language-based code generation components. To assess whether this modularity is necessary, we implement a unified configuration in which a single multimodal model performs both feedback interpretation and code synthesis. The unified variant shows substantially degraded performance, often producing invalid or incomplete reward code and failing to improve control behavior over iterations. These observations suggest that current vision–language models do not yet possess the compositional or programmatic reasoning required to perform both tasks simultaneously, underscoring the importance of maintaining distinct feedback and synthesis stages.

6. Conclusion and Discussion

In this work, we propose *MoVLR* , an automatic workflow that leverages vision-language models to bridge explicit language descriptions with the implicit motor control required for high-dimensional musculoskeletal systems. By incorporating multimodal feedback into the learning loop, our approach designs biomechanically grounded reward functions, which are iteratively refined to guide the musculoskeletal agent toward natural, stable motion. Through this method, we demonstrate that VLMs can successfully translate high-level motion descriptions into detailed control objectives, optimizing musculoskeletal agents' performance in diverse environments and tasks.

Experimental results across locomotion and manipulation tasks show that *MoVLR* outperforms human and language-based baselines, yielding greater task performances. The iterative refinement process allows the reward function to evolve from emphasizing coarse global stability to fine-grained joint coordination, closely reflecting the hierarchical organization of human motor learning.

Beyond quantitative improvements, *MoVLR* highlights a fundamental connection between *explicit language intent* and *implicit reward emergence*. The VLM acts as a perceptual intermediary that grounds linguistic objectives in physical dynamics, producing internal reward representations that are interpretable and dynamically consistent. This work opens possibilities for scaling biologically plausible control to more complex behaviors and morphologies. By leveraging multimodal, language-conditioned feedback, the proposed framework offers a promising path toward interpretable, adaptive, and generalizable control, where reward learning is shaped by perceptual understanding of motion and task intent.

References

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

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

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.

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.

Alberto Silvio Chiappa, Alessandro Marin Vargas, Ann Huang, and Alexander Mathis. Latent exploration for reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2023.

Yusen Feng, Xiyan Xu, and Libin Liu. Musclevae: Model-based controllers of muscle-actuated characters, 2023. URL https://arxiv.org/abs/2312.07340.

Yuying Ge, Annabella Macaluso, Li Erran Li, Ping Luo, and Xiaolong Wang. Policy adaptation from foundation model feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19059–19069, 2023.

Prasoon Goyal, Scott Niekum, and Raymond J Mooney. Using natural language for reward shaping in reinforcement learning. *arXiv preprint arXiv:1903.02020*, 2019.

Xingang Guo, Darioush Keivan, Usman Syed, Lianhui Qin, Huan Zhang, Geir Dullerud, Peter Seiler, and Bin Hu. Controlagent: Automating control system design via novel integration of llm agents and domain expertise. *arXiv preprint arXiv:2410.19811*, 2024.

Nicklas Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. *arXiv preprint arXiv:2310.16828*, 2023.

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.

Zhenyu Jiang, Yuqi Xie, Jinhan Li, Ye Yuan, Yifeng Zhu, and Yuke Zhu. Harmon: Whole-body motion generation of humanoid robots from language descriptions. *arXiv preprint arXiv:2410.12773*, 2024.

Seunghwan Lee, Moonseok Park, Kyoungmin Lee, and Jehee Lee. Scalable muscle-actuated human simulation and control. *ACM Transactions On Graphics (TOG)*, 38(4):1–13, 2019.

Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Eureka: Human-level reward design via coding large language models. *arXiv preprint arXiv:2310.12931*, 2023.

Yecheng Jason Ma, William Liang, Hungju Wang, Sam Wang, Yuke Zhu, Linxi Fan, Osbert Bastani, and Dinesh Jayaraman. Dreureka: Language model guided sim-to-real transfer. In *Robotics: Science and Systems (RSS)*, 2024.

Shinya Masadome and Taku Harada. Reward design using large language models for natural language explanation of reinforcement learning agent actions. *IEEJ Transactions on Electrical and Electronic Engineering*, 2025.

Mohammad Narimani and Seyyed Ali Emami. Agenticcontrol: An automated control design framework using large language models. *arXiv preprint arXiv:2506.19160*, 2025.

Jungnam Park, Sehee Min, Phil Sik Chang, Jaedong Lee, Moon Seok Park, and Jehee Lee. Generative gaitnet. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022.

Jungnam Park, Euikyun Jung, Jehee Lee, and Jungdam Won. Magnet: Muscle activation generation networks for diverse human movement. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–11, 2025.

Juan Rocamonde, Victoriano Montesinos, Elvis Nava, Ethan Perez, and David Lindner. Vision-language models are zero-shot reward models for reinforcement learning. *arXiv preprint arXiv:2310.12921*, 2023.

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.

Merkourios Simos, Alberto Silvio Chiappa, and Alexander Mathis. Reinforcement learning-based motion imitation for physiologically plausible musculoskeletal motor control. *arXiv preprint arXiv:2503.14637*, 2025.

Jiayang Song, Zhehua Zhou, Jiawei Liu, Chunrong Fang, Zhan Shu, and Lei Ma. Self-refined large language model as automated reward function designer for deep reinforcement learning in robotics. *arXiv preprint arXiv:2309.06687*, 2023.

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*, 18:1–17, 2021.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

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.

Seyed Shahabaldin Tohidi, Yildiray Yildiz, and Ilya Kolmanovsky. Fault tolerant control for over-actuated systems: An adaptive correction approach. In *2016 American control conference (ACC)*, pages 2530–2535. IEEE, 2016.

Yufei Wang, Zhanyi Sun, Jesse Zhang, Zhou Xian, Erdem Biyik, David Held, and Zackory Erickson. Rl-vlm-f: Reinforcement learning from vision language foundation model feedback. *arXiv preprint arXiv:2402.03681*, 2024.

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.

Tianbao Xie, Siheng Zhao, Chen Henry Wu, Yitao Liu, Qian Luo, Victor Zhong, Yanchao Yang, and Tao Yu. Text2reward: Automated dense reward function generation for reinforcement learning. In *International Conference on Learning Representations (ICLR), 2024 (07/05/2024-11/05/2024, Vienna, Austria)*, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Runhao Zeng, Dingjie Zhou, Qiwei Liang, Junlin Liu, Hui Li, Changxin Huang, Jianqiang Li, Xiping Hu, and Fuchun Sun. Video2reward: Generating reward function from videos for legged robot behavior learning. *arXiv preprint arXiv:2412.05515*, 2024.

Zhongchao Zhou, Yuxi Lu, Yaonan Zhu, Yifan Zhao, Bin He, Liang He, Wenwen Yu, and Yusuke Iwasawa. Llms-guided adaptive compensator: Bringing adaptivity to automatic control systems with large language models. *arXiv preprint arXiv:2507.20509*, 2025.

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 A.

A.1. Model Predictive Control with Morphology-Aware Proportional Control

Model Predictive Control with Morphology-Aware Proportional Control (MPC^2) (Wei et al., 2025) is a hierarchical control scheme for high-dimensional musculoskeletal systems. Let $z \in \mathbb{R}^{d_z}$ denote the major joint coordinates defining the system posture ($d_z \ll d_u$, where d_u is the actuator dimension). The high-level planner solves

$$z^* = \arg\min_z \sum_{h=0}^{H-1} C(s_{t+h}, u_{t+h}), \quad u_{t+h} = \pi_{\mathrm{MP}}(s_{t+h}, z), \tag{6}$$

using a sampling-based MPC (e.g., MPPI) over posture space. Instant rollouts are introduced by sampling candidate z around the current posture $M_{\mathrm{pos}}(s_t)$ for rapid recovery from disturbances.

The low-level morphology-aware proportional controller maps the target posture z^* to target actuator lengths l^* and computes desired actuator forces

$$f_m^* = \min\big(0,\, K \cdot (l^* - l)\big), \quad K = \bar{k} \sum_{i \in I_z} \big|\mathrm{col}_i(J_m) \cdot \big(z_i^* - M_{\mathrm{pos}}(s_t)_i\big)\big|\,, , \tag{7}$$

where K is the proportional gain of actuators, and $\bar{k}$ is the global scalar parameter. Actuator commands u^* follow from first-order actuator dynamics. This decomposition reduces the optimization dimension from $H \cdot d_u$ to d_z, enabling zero-shot control across morphologies without training.

A.2. Baseline Methods

Eureka (Ma et al., 2023) We adapt the Eureka framework, which uses large language models to synthesize reward functions from textual motion descriptions. For fair comparison, we implement Eureka using the same closed-loop setting as our method, but without the vision-language feedback: the language model receives textual summaries of agent rollouts rather than video-based feedback. The number of optimization rounds, samples per round, and other training parameters are matched to our method to ensure a controlled comparison.

HARMON (Jiang et al., 2024) We adapt the HARMON framework, which combines large language model reasoning with visual motion priors to generate whole-body humanoid motions. For fair comparison, we employ HARMON in our musculoskeletal control setting by using the same closed loop setting as our method, but replacing the video feedback with image feedback: the VLM receives 4 evenly spaced frames extracted from the rendered video rather than the full video. The number of optimization rounds, samples per round, and other training parameters are matched to our method to ensure a controlled comparison.

Human We use hand-crafted reward functions provided with the musculoskeletal tasks as a baseline. These rewards are designed by domain experts and encode task objectives through manually specified heuristics (Wei et al., 2025).

A.3. Prompts and Examples

Visual Feedback Prompt

```
You are given a video of a {body} muscle skeleton model whose task is to {task}.
     The video shows the model's actions after training a reinforcement learning
     model. The goal is to perform the task well and correctly. Provide a
    detailed critical analysis of the model's performance. You should be critial
     and point out specifically areas that need to be improved. Provide your
    analysis in a clear and concise manner, using appropriate technical language
     and terminology where necessary.

The reward terms used to train the reinforcement learning model, including their
     weights, are listed below.

{reward_terms}

Please perform the following analysis:

a) First determine whether the task {task} was completed successfully, answer
    YES or NO.
b) Identify the main issues with the motion produced compared to the desired
    motion from the task description. First focus on successfully completing the
     general task, then fine-tuning details. If the task is not successfully
    completed yet do not worry about fine-tuning details. Be detailed with
    descriptions. Also analyze what specific motions in the video could cause
    the issues or failures. Focus mainly on {focus} motion/issues. Describe
    directions from the point of the view of the muscle skeleton rather than a
    third person view.
c) Someone is trying to run a control method to perform better than what was
    shown in the video, and needs some suggestions about some reward terms that
    could be used, added, or given greater/less weight. For new terms, assign a
    reasonable weight value between 0 and the maximum weight, and increase/
    decrease gradually if/when necessary. Do not suggest too many or redundant
    terms. If suggesting a new term, also suggest how the function should be
    defined (using words is enough, don't need to use specific functions/coding
    names). Given the video, issues, and existing reward terms listed above,
    provide some suggestions.

Be specific in your observations and suggestions. Your goal is to help improve
    both the correctness and the naturalness of the {body}'s motion
```

Coding Language Model Prompt

```
You are updating the residual function of a MuJoCo muscle-skeleton environment
    using a **conservative, feedback-driven edit policy** to improve the
    performance of the task.

**Inputs**
- Goal/task: {task}
- Environment code (contains residual function):
$[$env_code$]$
{env_code}
$[$env_code$]$
- Task file (canonical list of valid sensors):
$[$task_code$]$
{task_code}
$[$task_code$]$
- The weights for each residual term during previous stages are provided below.
```

```
{residual_terms}
- Video feedback after analyzing a single round of running the muscle skeleton
    performing the task:
{feedback_string}

**Editing guidelines**
- Make a **small number of localized changes** that directly address issues
    observed in the feedback.
- When possible, prefer adjusting existing residual terms (e.g., scaling,
    weighting, or tuning) before introducing new ones.
- New residual terms may be added if they clearly align with the feedback and
    are supported by the task file.
- Residual functions should be defined carefully and with enough detail
- Keep edits focused on the relevant regions; avoid broad or unrelated
    modifications.
- Do not include weight term implementations in the environment code, all terms
    should be multiplied by 1.
- Pay attention to comments in the code if they exist in the code
- Ensure that the residual function remains stable and interpretable across
    training stages.
- When editing the residual function, the following vector/quaternion operations
     can be used

{operations}

{code_tips}

**Output format (strict)**
- Output the **entire, updated environment code** in a single ```cpp``` block.
- No explanations, no diffs, no comments, only the final code.
```

Selection

```
You are an expert biomechanical analyst. You will be shown two videos, each
    depicting a muscle-skeleton model performing the task {task}. Carefully
    observe both performances and compare how accurately, smoothly, and
    efficiently the models complete the task.

Evaluate each video based on key biomechanical factors: task success, balance
    and stability, posture and alignment, joint coordination, and overall
    movement naturalness. Consider whether the motion looks physically plausible
     and efficient, without unnecessary or unstable compensations. Pay attention
     to gait or limb symmetry, center-of-mass control, and the sequencing of
    major joints.

After analyzing both videos, choose which one demonstrates better completion of
    the task, that is, which looks more correct, natural, stable, and
    biomechanically efficient.
Respond with only one of the following words: "first" or "second", followed by a
     brief explanation justifying your choice.
```

Task Descriptions

Flat Terrain

```
Walk forward in a straight line on flat terrain at a velocity of about 1 m/s,
    maintaining directional stability and good posture
```

Slope Terrain

```
Walk forward in a straight line on sloped terrain at a velocity of about 1 m/s,
    maintaining directional stability and good posture
```

Rough Terrain

```
Walk forward in a straight line on rough terrain at a velocity of about 1 m/s,
    maintaining directional stability and good posture
```

Bottle Pouring

```
Grasp and reorient the darker-shaded bottle to match the target orientation and
    position, indicated by a lighter shade bottle
```

Cube Rotation

```
Grasp and reorient the cube to match the target orientation, keeping the cube
    held approximately in front of the musculoskeleton's chest
```

Ostrich

```
Make an ostrich walk forward in a straight line on a flat terrain with velocity
    approximately 1 m/s and proper gait and posture (flat body, relatively
    straight legs, stable head)
```

Injured Body

```
Make a human muscle skeleton model with right-side injuries to the biceps,
    gastrocnemius, semimembranosus, and semitendinosus muscles walk forward in a
     straight line on a flat terrain
```

Left-Turn

```
Walk forward with good posture, then make a left turn and walk towards the new
    facing direction after making the turn
```

A.4. Additional Experimental Results

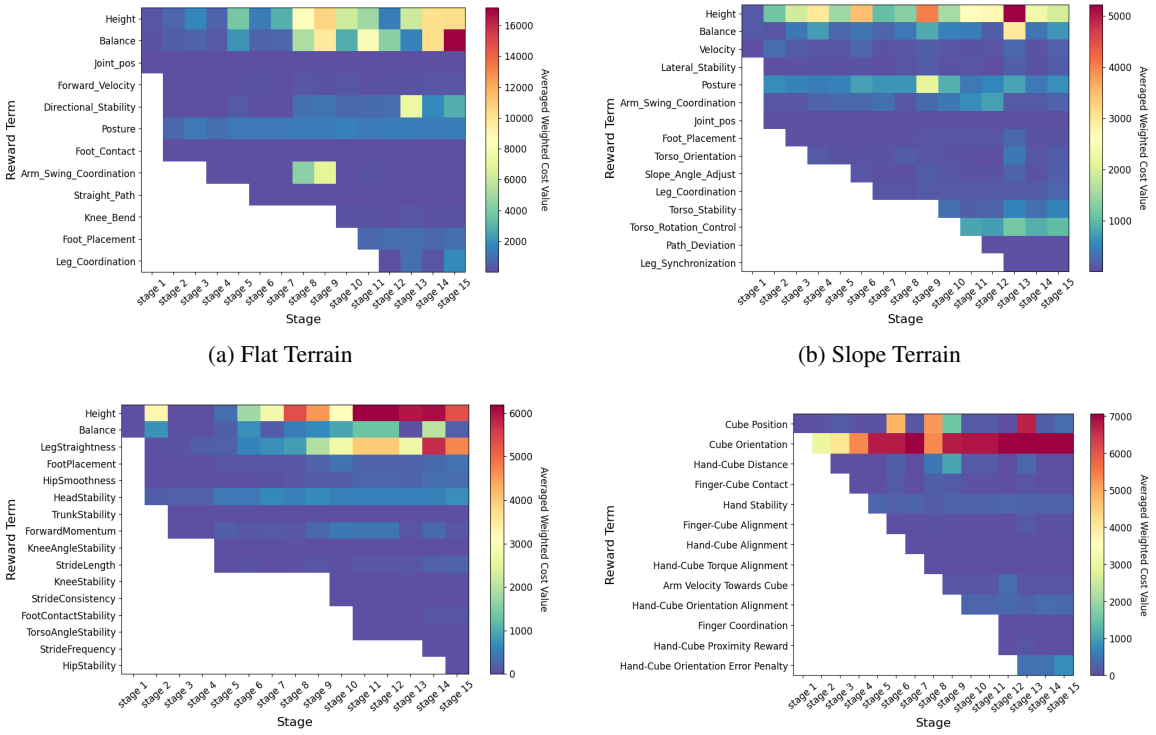
Evolution of weighted reward terms for remaining tasks

We provide supplementary heatmaps visualizing the evolution of reward term weights across refinement stages for the remaining four tasks: flat terrain, slope terrain, ostrich locomotion, and cube rotation.

Comparison of Language-Designed and Human-Defined Reward Terms

Figure 8 compares the residual reward terms designed by the language model with those manually specified by human experts across three musculoskeletal systems. The comparison highlights the model's capacity to infer a more comprehensive and morphology-aware set of control objectives.

In the fullbody model, the language model introduces a broader range of biomechanically grounded terms – such as *pelvis tilt control, hip coordination*, and *fgait symmetry* – which extend beyond the coarse global stability terms (*height, velocity, balance*) typically defined by human experts. For the upperbody model, the model captures fine-grained kinematic relations including *elbow strength, wrist rotation*, and per-finger coordination, reflecting an understanding of localized control relevant to manipulation tasks. Finally in the ostrich model, the model adapts to non-human morphology with terms such as *neck height, torso angle*, and *head stability*, indicating a morphological generalization beyond human-centered priors.



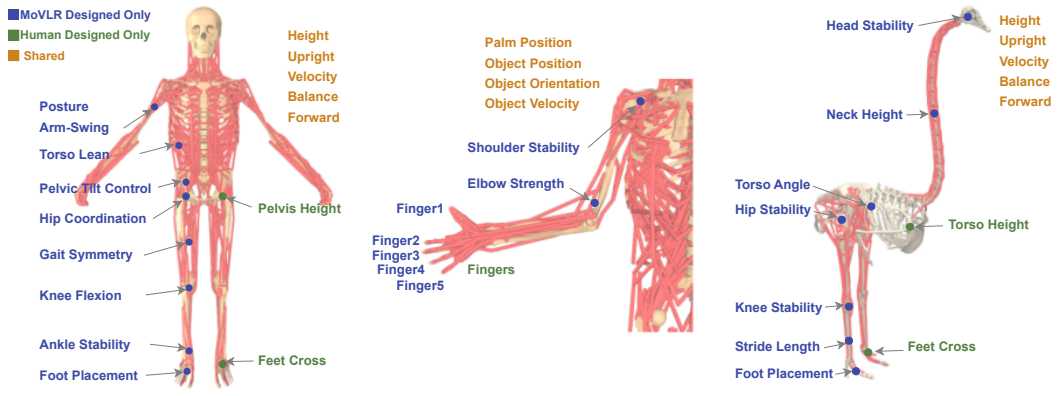
(a) Flat Terrain

(b) Slope Terrain

(c) Ostrich

(d) Cube Rotation

Figure 7: Weighted reward terms for (a) flat terrain, (b) slope terrain, (c) ostrich, (d) cube rotation



a) Fullbody MS Model

b) Upperbody MS Model

c) Ostrich MS Model

Figure 8: Comparison of reward terms desgined by LLM only (blue) and by human experts (green), with shared terms shown in orange, across three musculoskeletal systems.