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*

Editors: G. Sukhatme, L. Lindemann, S. Tu, A. Wierman, N. Atanasov

Abstract

Discovering effective reward functions remains a fundamental challenge in motor control of high-dimensional musculoskeletal systems. While humans can describe movement goals explicitly such as "walking forward with an upright posture," the underlying control strategies that realize these goals are largely implicit, making it difficult to directly design rewards from high-level goals and natural language descriptions. We introduce Motion from Vision-Language Representation (*MoVLR*), a framework that leverages vision-language models (VLMs) to bridge the gap between goal specification and movement control. Rather than relying on handcrafted rewards, *MoVLR* iteratively explores the reward space through iterative interaction between control optimization and VLM feedback, aligning control policies with physically coordinated behaviors. Our approach transforms language and vision-based assessments into structured guidance for embodied learning, enabling the discovery and refinement of reward functions for high-dimensional musculoskeletal locomotion and manipulation. These results suggest that VLMs can effectively ground abstract motion descriptions in the implicit principles governing physiological motor control.

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

1. Introduction

Humans acquire motor 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, where multiple muscles or synergies can produce identical joint motions. Coordinated movement therefore depends on discovering appropriate patterns of whole-body control rather than specifying individual actuator commands, making the realization of efficient and natural motion fundamentally challenging.

Learning-based approaches have enabled progress in high-dimensional musculoskeletal control (Caggiano et al., 2022; Schumacher et al., 2023; He et al., 2024). However, most existing methods rely on heuristic objectives such as velocity tracking or energy minimization, which often fail to capture the nuanced structure of motion complexity. While sufficient for basic task completion, such objectives often neglect anatomical principles and lead to biomechanically unnatural or inefficient behaviors.

Recent advances in large language models (LLMs) and vision-language models (VLMs) suggest a promising direction for exploring reward structure through high-level assessments of motion quality and coordination (Sontakke et al., 2023; Ma et al., 2024a; Zeng et al., 2024). Nevertheless, existing approaches typically rely on episodic statistics or coarse success signals as feedback and are evaluated primarily on low-dimensional, torque-driven systems with explicit dynamics. Motor control, however, is governed by temporally extended and implicitly structured sensorimotor dynamics that are difficult to capture through such feedback alone. As a result, it remains unclear whether these models can systematically translate implicit interaction dynamics into structured rewards for high-dimensional musculoskeletal control.

In this paper, we introduce Motion from Vision Language Representation (*MoVLR*), a framework for automatic reward learning in high-dimensional musculoskeletal systems that integrates descriptive and dynamical feedback. As shown in Figure 1, *MoVLR* extracts musculoskeletal dynamics through policy optimization and rollout, rendering the resulting behaviors as movement videos. A vision-language model evaluates these motions to produce structured biomechanical feedback, which is then used by a language model to refine the reward formulation in subsequent iterations. Through this iterative process, implicit dynamical regularities are progressively distilled into explicit reward terms. By grounding temporal dynamics in semantically meaningful descriptors, *MoVLR* bridges perceptual feedback with domain-informed motion representations, enabling scalable and biomechanically realistic reward discovery for high-dimensional musculoskeletal control. The supplementary material and experiment demonstrations can be found at: https://lnsgroup.cc/research/MoVLR/.

Contributions: (1) We present *MoVLR*, a fully automatic framework for discovering reward structure in high-dimensional musculoskeletal systems by capturing implicit dynamics and refining them into explicit, executable rewards for control. (2) We demonstrate that *MoVLR* generalizes across movement tasks, environments, and morphologies, producing interpretable reward terms that reflect underlying musculoskeletal dynamics. (3) We introduce a unified framework that enables interpretable evaluation of motor performance, adaptive reward refinement through vision–language feedback, and transferable control for natural and coordinated motion.

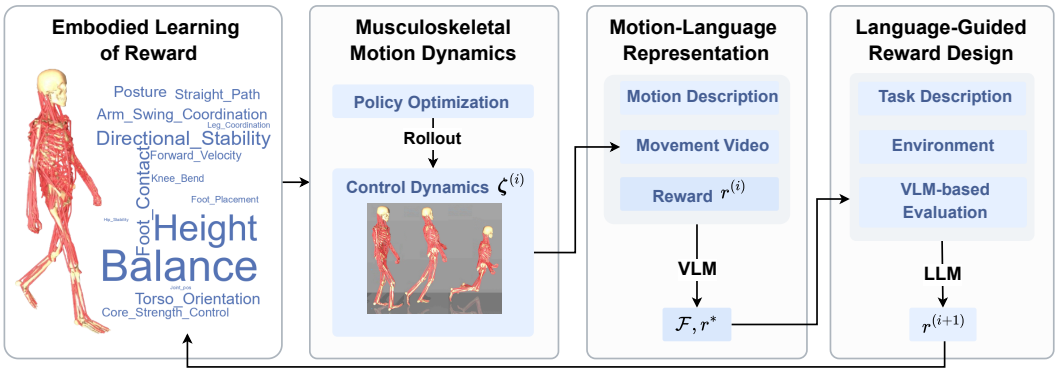


Figure 1: Workflow of *MoVLR*. Policy optimization is performed to provide high-dimensional musculoskeletal dynamics of the reward candidate. A VLM evaluates the corresponding movement video $\zeta^{(i)}$ to update the current best reward design r^* and suggest biomechanical improvements $\mathcal{F}$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

2. Related Works

2.1. Control of musculoskeletal systems

The control of high-dimensional musculoskeletal systems remains challenging due to the redundancy and dimensionality of human-like actuation. Significant progress has been made in developing faithful simulation environments that model 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 address control complexity, prior work has explored hierarchical decompositions (Lee et al., 2019; Park et al., 2022; Feng et al., 2023), curriculum learning strategies (Caggiano et al., 2023; Park et al., 2025), bio-inspired sampling (Schumacher et al., 2023), latent-space coordination (Chiappa et al., 2023), and model-based control (Hansen et al., 2024). More recent approaches further reduce effective dimensionality by extracting muscle synergies informed by anatomy or task structure (Berg et al., 2023; He et al., 2024).

Despite these advances, achieving natural and human-like behaviors remains difficult. Performance is highly sensitive to reward design, where small changes can lead to unnatural postures, inefficient coordination, or brittle behaviors. Crafting such rewards typically requires domain expertise and manual tuning, and often relies on indirect proxies rather than direct biomechanical measures of movement quality (Song et al., 2021; Sui et al., 2017). This reliance on hand-crafted objectives motivates the development of more expressive and adaptive reward design mechanisms, naturally leading to language and multimodal-driven approaches.

2.2. Language and multimodal driven reward design

Recent advances in large language models have shown strong potential for facilitating reward and feedback design in robotics and simulation systems (Goyal et al., 2019; Ma et al., 2024b; Yu et al., 2023). Eureka (Ma et al., 2024a), for example, uses code-generating LLMs to synthesize dense reward functions that exceed manually engineered counterparts in expressivity and task relevance. Complementary to this, language-conditioned reward modulation has also been explored as a mechanism for shaping policy learning during pretraining, enabling more flexible and semantically aligned optimization objectives (Adeniji et al., 2024). Although originally introduced for reinforcement learning, this paradigm is equally applicable to robotics control, where generated signals can be interpreted as structured feedback shaping system dynamics toward desired trajectories (Brohan et al., 2023; Driess et al., 2023).

Beyond text-only models, recent work with vision–language models demonstrates the benefits of incorporating multimodal inputs such as images or video into feedback design (Rocamonde et al., 2024; Zitkovich et al., 2023; Ge et al., 2023; Wang et al., 2024; Zeng et al., 2024). For example, HARMON (Jiang et al., 2025) leverages a VLM to iteratively refine humanoid motion by evaluating rendered frame sequences against language descriptions. More recent work uses VLMs to detect and reason about failure modes in robotic manipulation, enabling richer evaluative feedback to inform reward design (Duan et al., 2025). Despite these advances, existing approaches lack a principled framework for transforming LLM and VLM feedback into structured dynamical signals that directly shape reward functions, instead relying on heuristic or loosely coupled representations.

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 formalize 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:

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

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 π_r^* is the optimal policy derived by maximizing r. In practice, the global reward R is specified at a high level using natural language descriptions such as "walk forward" or "grasp 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. Effective reward learning requires two key components: (1) **Embodied understanding of human movement** which extracts implicit biomechanical knowledge from motion descriptions, and (2) **Effective reward design** which integrates multimodal feedback to produce interpretable and executable reward terms 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 requiring hours to days of training, MPC^2 employs

a training-free pipeline which 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 specification.

4.2. Motion-language representation

We use 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 definition. If the newly proposed reward yields control sequences that better align with the motion description, both the current best reward r^* and the corresponding control dynamics $\boldsymbol{\zeta}^*$ 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 synthesis. 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.

In practice, to improve robustness, motion evaluations are verified through multiple VLM assessments to reduce the impact of inconsistent feedback. When incorrect evaluations occur, resulting reward updates typically fail to improve performance and are rejected in subsequent iterations, ensuring unreliable feedback does not propagate through the refinement process.

4.3. Language-guided reward design

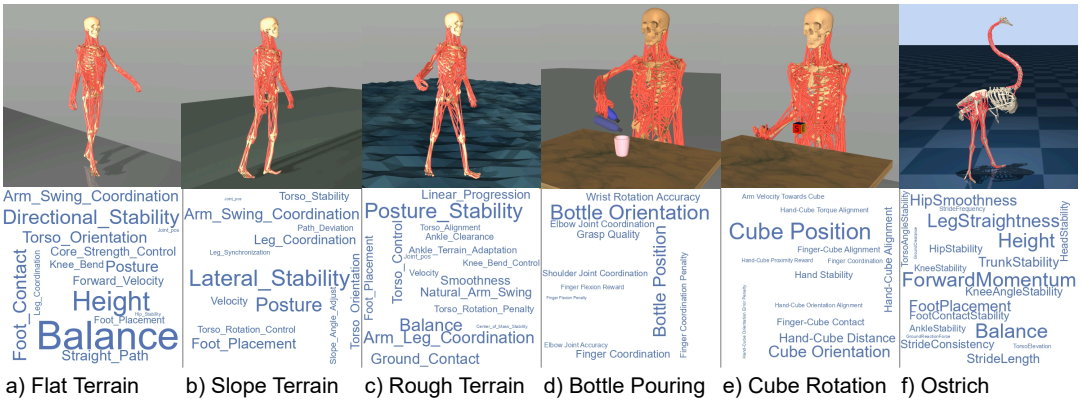
Designing executable reward functions from multimodal inputs requires mapping semantic and structural priors to physically meaningful quantities. In musculoskeletal control, this must capture nonlinear couplings between balance, posture, and coordination which are difficult to encode manually. We employ 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, and 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 the contextual information and VLM-based evaluations $\mathcal{F}$, the LLM performs a local search over the current best reward function r^* to synthesize new reward terms. Each term encodes a biomechanical sub-objective such as orientation tracking, smoothness, or joint stability. At each iteration, candidate rewards are evaluated through control rollouts and accepted only if they improve performance. The inherent non-determinism of LLM and VLM outputs encourages exploration across iterations, preventing the search from getting stuck in suboptimal solutions even when proposals are rejected. This process continues until an effective reward function is identified, yielding an interpretable objective suitable for policy optimization and control-based dynamical feedback.

Compared with traditional language-based approaches that rely solely on LLMs, *MoVLR* incorporates dynamical control feedback, enabling the VLM to reflect on kinematic and postural precision, an essential capability for achieving stable musculoskeletal control.



a) Flat Terrain b) Slope Terrain c) Rough Terrain d) Bottle Pouring e) Cube Rotation f) Ostrich

Figure 2: 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 explore the space of possible reward functions, solve novel tasks, and integrate different forms of human input. Unless otherwise noted, all VLM and LLM-based reward 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.

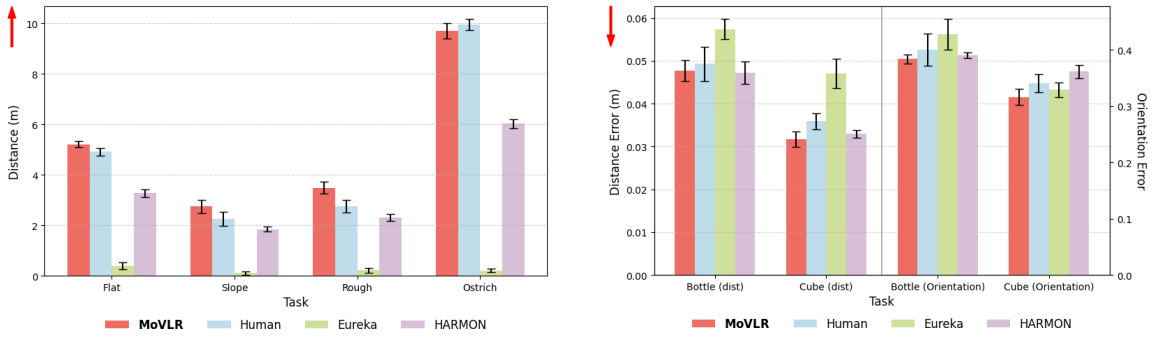
5.1. Environments

As shown in Figure 2, our experimental setup comprises a diverse set of musculoskeletal systems and tasks in the MuJoCo simulator (Todorov et al., 2012), capturing a broad spectrum of control challenges. The suite includes locomotion environments spanning flat, rough, and sloped terrain, testing stability and adaptability under varying conditions. Within the flat-terrain setting, we further consider directional turning and an injured-body condition where selected leg muscle groups are weakened, enabling evaluation of gait robustness and compensatory strategies. In addition to locomotion, the setup includes manipulation tasks such as bottle pouring and cube manipulation that emphasize coordination and precision, as well as a non-human locomotion task based on an ostrich muscle model to assess 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., 2024a), and HARMON (Jiang et al., 2025) on both locomotion and manipulation tasks. Our implementation details of method and baselines are elaborated in Appendix A.2. Evaluation focuses on final performance after convergence, measured by average walking distance over 10 seconds for locomotion and by object position and orientation errors for manipulation.

As shown in Figure 3 (a), across all locomotion environments, *MoVLR* consistently achieves higher task performance, yielding the longest walking distances on flat, sloped, and rough terrains. The gains are most pronounced in challenging settings, where terrain irregularities demand adaptive stability and coordinated motion. Compared to HARMON, which relies on visual alignment over



(a) Locomotion Tasks (b) Manipulation Tasks

Figure 3: 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).

discrete frame sequences, *MoVLR* integrates feedback more tightly with control dynamics by using full trajectory rollouts to inform structured reward updates, enabling the capture of temporally extended dependencies such as gait rhythm, balance recovery, and joint coordination. Although *MoVLR* performs slightly below the human baseline, it generalizes effectively to the ostrich environment, maintaining strong performance despite significant morphological differences. These results indicate that multimodal reward refinement produces robust and transferable control objectives across biomechanical structures.

As shown in Figure 3 (b), *MoVLR* achieves the lowest average position and orientation errors compared to all baselines in the 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.

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 4, 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 5 (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 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 reg-

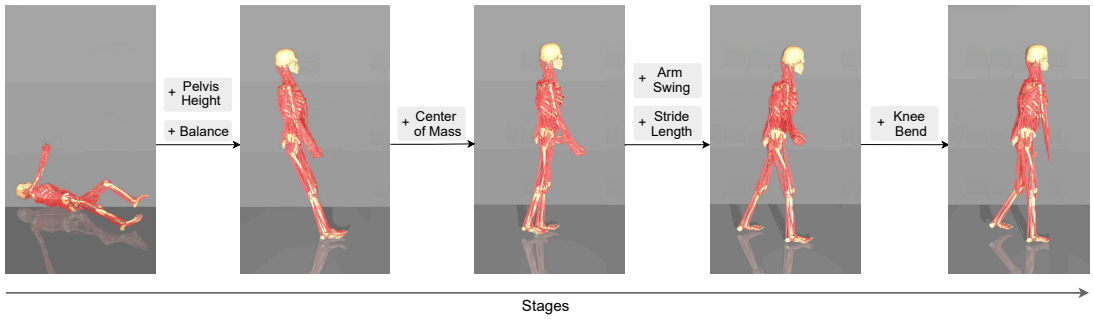


Figure 4: Progressive improvement of the musculoskeletal model's gait across training stages based on movement video.

ulation. 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.

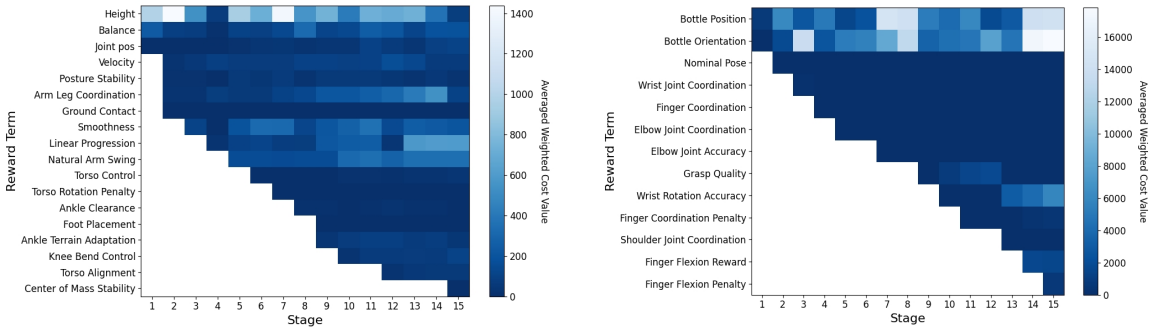
In the bottle pouring task shown in Figure 5 (b), a similar hierarchical refinement pattern is observed. Initial stages emphasize gross spatial alignment through terms such as *bottle position* and *bottle orientation*, enabling task feasibility. With continued refinement, weights shift toward fine motor 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.

Additional heatmaps for the remaining four tasks are included in appendix A.5, illustrating consistent refinement dynamics across both locomotion and manipulation domains.

Reward Interpretability. The learned reward terms exhibit strong semantic and structural interpretability, with many feature names corresponding to well-defined biomechanical quantities that admit explicit formulations. For example, *height* is often expressed as $h = z_{\text{COM}}$, the vertical position of the center of mass. Similarly *forward velocity* can be represented as $v_x = \dot{x} * \text{COM}$, and *balance* as the deviation between the projected center of mass and the support polygon, e.g., $|\text{IIsupport}(x * \text{COM}) - x_{\text{foot}}|$, etc. More generally, reward terms can be viewed as encodings of desired biomechanical properties, where these formulations provide a direct pathway to constructing executable reward functions that can be optimized within the control framework.

Ablation Studies. To better understand the contribution of each 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.

Reward generalizability. We test transferring reward functions proposed 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.



(a) Rough Terrain

(b) Bottle Pouring

Figure 5: Weighted reward terms per stage for (a) locomotion task, (b) manipulation task

We further 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.

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, highlighting the importance of maintaining distinct feedback and synthesis stages.

6. Conclusion and Discussion

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

Experimental results across locomotion and manipulation tasks show that *MoVLR* consistently outperforms both human-designed and language-based baselines. The iterative refinement of rewards mirrors the hierarchical structure of human motor learning, progressing from coarse stability constraints to fine-grained joint coordination.

Beyond performance gains, *MoVLR* highlights a fundamental connection between *explicit language intent* and *implicit reward emergence*. Acting as a perceptual bridge, the VLM grounds linguistic goals in physical dynamics, producing interpretable and dynamically consistent reward representations. This work offers a scalable and principled path toward biologically plausible, interpretable, and generalizable control for complex behaviors and morphologies.

Acknowledgments

This work is supported by STI 2030-Major Projects 2022ZD0209400 and NSFC 62461160313. Correspondence to: Yanan Sui (ysui@tsinghua.edu.cn).

References

Ademi Adeniji, Amber Xie, Carmelo Sferrazza, Younggyo Seo, Stephen James, and Pieter Abbeel. Language reward modulation for pretraining reinforcement learning. In *Workshop on Training Agents with Foundation Models at RLC*, 2024. URL https://arxiv.org/abs/2308.12270.

Cameron H. Berg, Vittorio Caggiano, and Vikash Kumar. Sar: Generalization of physiological agility and dexterity via synergistic action representation. In *Proceedings of Robotics: Science and Systems*, page 7, 2023.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, et al. Rt-1: Robotics transformer for real-world control at scale. In *Proceedings of Robotics: Science and Systems*, 2023.

Vittorio Caggiano, Huawei Wang, Guillaume Durandau, Massimo Sartori, and Vikash Kumar. Myosuite: A contact-rich simulation suite for musculoskeletal motor control. In *Proceedings of The 4th Annual Learning for Dynamics and Control Conference*, pages 492–507. PMLR, 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.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, et al. Palm-e: An embodied multimodal language model. *CoRR*, abs/2303.03378, 2023.

Jiafei Duan, Wilbert Pumacay, Nishanth Kumar, Yi Ru Wang, Shulin Tian, Wentao Yuan, Ranjay Krishna, Dieter Fox, Ajay Mandlekar, and Yijie Guo. Aha: A vision-language model for detecting and reasoning over failures in robotic manipulation. In *Proceedings of the Thirteenth International Conference on Learning Representations*, 2025.

Yusen Feng, Xiyan Xu, and Libin Liu. Musclevae: Model-based controllers of muscle-actuated characters. In *SIGGRAPH Asia 2023 Conference Papers*, SA '23. Association for Computing Machinery, 2023.

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. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 2385–2391, 2019.

Nicklas Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. In *Proceedings of The Twelfth International Conference on Learning Representations*, 2024.

Kaibo He, Chenhui Zuo, Chengtian Ma, and Yanan Sui. DynSyn: Dynamical synergistic representation for efficient learning and control in overactuated embodied systems. In *Proceedings of the 41st International Conference on Machine Learning*, pages 18115–18132. PMLR, 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. In *Proceedings of The 8th Conference on Robot Learning*, pages 3015–3026. PMLR, 2025.

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

Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Jim Fan, and Anima Anandkumar. Eureka: Human-level reward design via coding large language models. In *International Conference on Representation Learning*, pages 26516–26560, 2024a.

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*, 2024b.

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. In *The Twelfth International Conference on Learning Representations*, 2024.

Pierre Schumacher, Daniel F.B. Haeufle, Dieter Büchler, Syn Schmitt, and Georg Martius. Dep-rl: Embodied exploration for reinforcement learning in overactuated and musculoskeletal systems. In *Proceedings of The Eleventh International Conference on Learning Representations*, 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.

Sumedh Sontakke, Jesse Zhang, Séb Arnold, Karl Pertsch, Erdem Bıyık, Dorsa Sadigh, Chelsea Finn, and Laurent Itti. Roboclip: One demonstration is enough to learn robot policies. *Advances in Neural Information Processing Systems*, 36:55681–55693, 2023.

Yanan Sui, Kun ho Kim, and Joel W Burdick. Quantifying performance of bipedal standing with multi-channel emg. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3891–3896. IEEE, 2017.

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.

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. In *Proceedings of the 41st International Conference on Machine Learning*, pages 51484–51501. PMLR, 2024.

Yunyue Wei, Shanning Zhuang, Vincent Zhuang, and Yanan Sui. Motion control of high-dimensional musculoskeletal systems with hierarchical model-based planning. In *The Thirteenth International Conference on Learning Representations*, 2025.

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.

Wenhao Yu, Nimrod Gileadi, Chuyuan Fu, Sean Kirmani, Kuang-Huei Lee, Montserrat Gonzalez Arenas, Hao-Tien Lewis Chiang, Tom Erez, Leonard Hasenclever, Jan Humplik, Brian Ichter, Ted Xiao, Peng Xu, Andy Zeng, Tingnan Zhang, Nicolas Heess, Dorsa Sadigh, Jie Tan, Yuval Tassa, and Fei Xia. Language to rewards for robotic skill synthesis. In *Proceedings of The 7th Conference on Robot Learning*, pages 374–404. PMLR, 2023.

Runhao Zeng, Dingjie Zhou, Qiwei Liang, Junlin Liu, Hui Li, Changxin Huang, Jianqiang Li, Xiping Hu, and Fuchun Sun. Video2reward: Generating reward functions from videos for legged robot behavior learning. In *Proceedings of the 27th European Conference on Artificial Intelligence*. IOS Press, 2024.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of The 7th Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.

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*, 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., 2024a) 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., 2025) 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. Additional discussion on Runtime, Inference Cost, and Scalability

We provide a discussion on the computational characteristics of our method in terms of runtime, inference cost, and scalability. A typical run of the framework involves multiple iterative stages of policy optimization, video rendering, and VLM/LLM inference. In our experiments, we use up to

15 stages, although strong performance can often be achieved within 5-10 stages. Per stage, policy optimization requires approximately 5–6 minutes, video rendering 1 minute, and each VLM/LLM inference 1 minute, yielding a total of 10 minutes per stage. This corresponds to 35–40 minutes for a 5-stage run and 105–120 minutes for a full 15-stage run. In terms of inference cost, each run consumes approximately 30,000 tokens (combined video and language input), with the exact monetary cost depending on the specific VLM/LLM used. Overall, the framework scales linearly with the number of stages and can be adjusted to trade off runtime and performance.

A.4. Prompts and Examples

Throughout our pipeline, we use multiple prompts to guide both perception and reward synthesis. The primary prompts correspond to (i) motion–language representation, where the VLM is prompted to evaluate rendered motion trajectories against a task description and produce structured feedback, and (ii) language-guided reward design, where the LLM is prompted to refine reward functions based on the task specification, environment context, and VLM feedback. Brief examples of these prompts are illustrated in Figure 6, while detailed, concrete prompt instances used in our experiments are provided below.

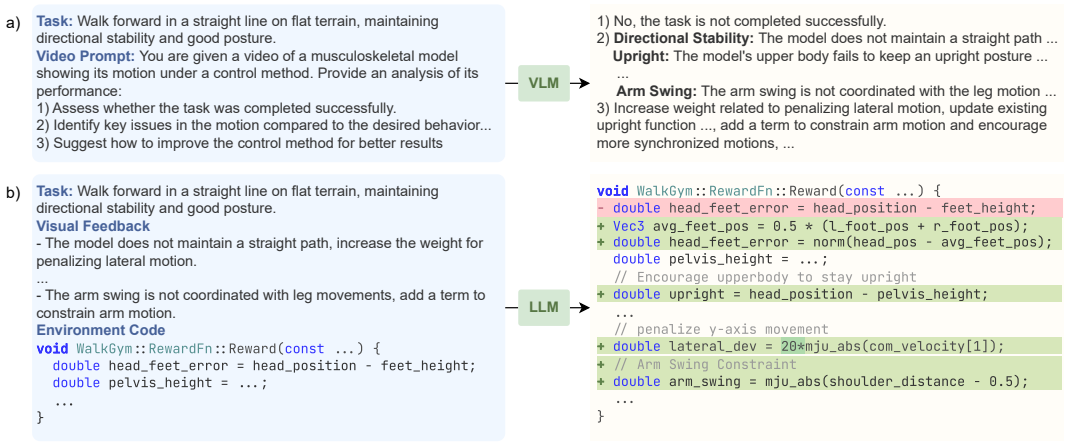


Figure 6: Example inputs and outputs of the (a) VLM, and (b) LLM. The VLM analyzes a video motion sequence based on the given motion description and provides diagnostic feedback. The LLM uses this feedback to explore corresponding code modifications to the reward function.

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
```

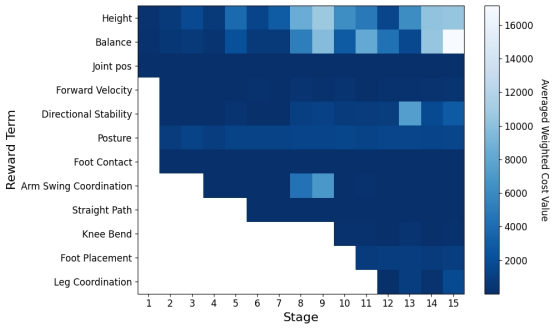
Left-Turn

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

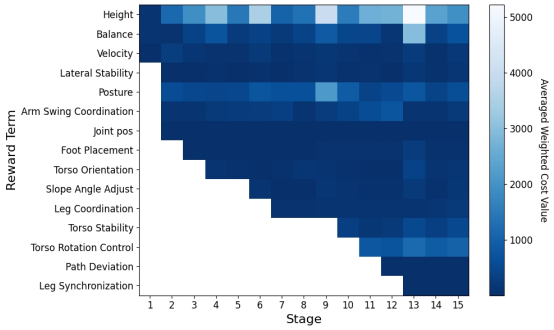
A.5. Additional Experimental Results

Evolution of weighted reward terms for remaining tasks

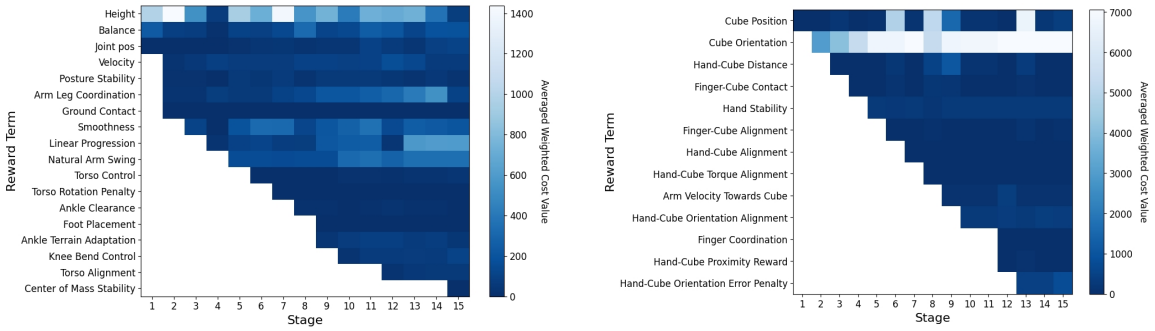
We provide supplementary heatmaps visualizing the evolution of reward term weights across refinement stages for the remaining four tasks: flat terrain, slope terrain, ostrich locomotion, and cube rotation.



(a) Flat Terrain



(b) Slope Terrain



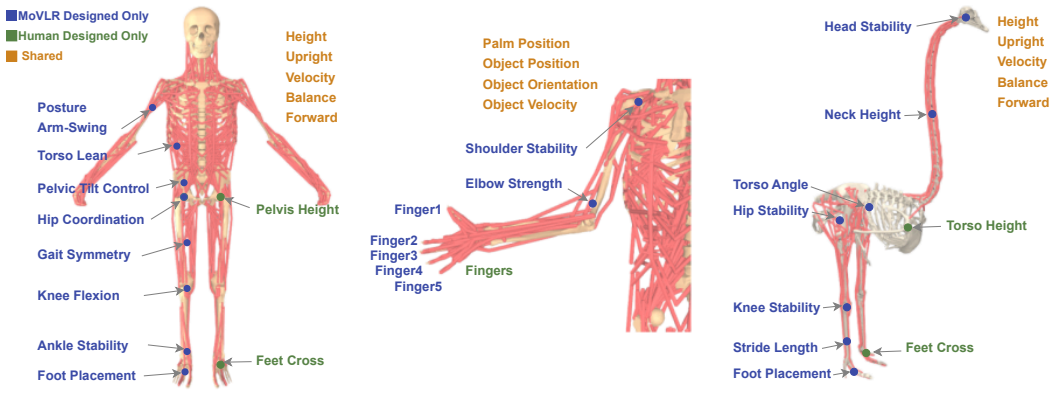
(c) Ostrich (d) Cube Rotation

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

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 *gait 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) Fullbody MS Model b) Upperbody MS Model c) Ostrich MS Model

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