2026 European Conference on Computing in Construction
Corfu, Greece
July 12–15, 2026

PHYSICS-CONSTRAINED REINFORCEMENT LEARNING FOR INDIVIDUALIZED AND BIOMECHANICALLY FEASIBLE TASK-MOTION OPTIMIZATION

Kangrui Ren[1], Ali Golabchi[2], and Gaang Lee[1]
[1]University of Alberta, Canada
[2]Department of Mechanical Engineering, University of Alberta, Canada

Abstract

The construction industry faces high musculoskeletal disorder rates due to repetitive, awkward task-postures. Current manual corrections and static optimizations often ignore individual anthropometry and whole-body biomechanical interdependencies, leading to physically unfeasible results. This study explores Reinforcement Learning (RL) for tailored posture optimization by developing a pipeline that overcomes technical barriers like reward instability and high computational costs. The technique predicts joint-space residuals via a distilled, risk-aware RL policy within a physics simulator. Testing in construction-like scenarios showed significantly reduced mean risk and high-risk spikes while maintaining task fidelity. These findings demonstrate that RL can provide personalized, executable ergonomic interventions.

Introduction

The construction industry continues to experience a high prevalence of work-related musculoskeletal disorders (WMSDs), imposing a substantial burden on both worker health and operational productivity (Santos et al., 2025). These disorders are rarely the result of a single event; rather, they accumulate through sustained or repetitive exposure to risk factors such as awkward postures, forceful exertions, and repetitive motions during task performance (Hulshof et al., 2021). Since these contributing factors typically emerge from the complex interaction between poor work design and a worker's habitual postural patterns, effective mitigation requires not only work-redesign interventions but also, critically, systematic correction of workers' working postures during task performance.

Current industrial practices for assessing and correcting task postures rely heavily on manual processes. Assessment is typically conducted using self-reports or observational tools such as the Rapid Upper Limb Assessment (RULA) (McAtamney and Corlett, 1993) or the Rapid Entire Body Assessment (REBA) (Hignett and McAtamney, 2000). In practice, correction strategies are often limited to verbal, on-the-spot instructions delivered by safety managers. While intuitive, these local corrections often fail to account for whole-body biomechanical interdependencies; correcting a local issue (e.g., lowering an elbow) may inadvertently create or worsen risk elsewhere (e.g., increasing lower back torque) due to inherent human biomechanical coupling and task constraints (Belbeck et al., 2014; Nadon et al., 2018; Rören et al., 2024). Because these interventions depend on generic ergonomic guidelines, they often fail to consider individual anthropometric differences and habitual motion patterns, thereby making it difficult to provide posture corrections that are truly personalized and readily adoptable (Ji et al., 2023; Chen and Zhang, 2025). These limitations underscore a need for a systematic technique to generate biomechanically coherent, coordinated whole-body posture corrections that are customized to the individual's specific anthropometric characteristics within the given task context.

To derive whole-body-coordinated posture corrections, researchers have applied mathematical optimization, such as trajectory optimization (optimal control) (Dafarra et al., 2022), and inverse optimization (Westermann et al., 2020). However, these existing approaches often simplify whole-body risk as an additive sum of per-joint or angle-based scores in their equations, which limits their ability to account for the coupled constraints across the body (González-Alonso et al., 2025). By failing to enforce these coupled constraints, such approaches overlook critical factors such as load transfer, support stability, and contact maintenance (Zheng and Yumane, 2015), often resulting in "optimizations" that are non-actionable or physically impossible. Furthermore, on the optimization algorithm side, current approaches often rely on static, frame-based reasoning that does not account for individual anthropometric variabilities or the temporal flow of motion (Kee, 2022; Ciccarelli et al., 2026), prioritizing a raw "risk" score reduction without considering feasibility, motion smoothness, or workflow continuity.

Reinforcement learning (RL) offers a promising alternative to overcome these limitations by addressing the complexities of human motion that static optimization misses. Unlike frame-based techniques, RL is uniquely positioned to learn full-body coordinated actions, allowing it to discover posture adjustments that remain biomechanically feasible at the global level. What frame-based techniques miss is temporal coupling: feasibility constraints (stability, contact evolution, inter-joint coordination) are trajectory-level, not per-frame. It can also condition learned policies on individual user attributes, enabling the generation of personalized motion strategies rather than generic templates. This is because RL optimizes expected cumulative reward over time, not a single-step objective: the agent

is trained on long-horizon rollouts where feasibility violations and risk spikes occur downstream, enabling credit assignments that link later outcomes back to earlier posture choices. With a state representation that includes task context and user attributes, the learned policy becomes both temporally aware (trajectory-level) and personalized (user-conditioned) in a single decision-making loop. Crucially, RL handles sequential, history-dependent decision-making (Mnih et al., 2015; Schulman et al., 2015), which allows for the optimization of entire motion trajectories rather than isolated frames. By encoding multiple objectives within the reward function, RL agents can learn balanced posture strategies that reflect necessary real-world trade-offs between ergonomic risk, task efficiency, and motion smoothness.

Despite this theoretical promise, RL has not yet been sufficiently tested or validated as a technique for individually tailored task-posture optimization due to several technical barriers. A primary challenge is risk signal reliability and reward stability; MSD risk signals and video-based skeletal data can be noisy or non-smooth, often causing unstable learning processes or the development of "shortcut" behaviors that game the reward system without truly reducing risk (Ren, 2025). Furthermore, ensuring executability and safety remains difficult, as safety is often treated via soft penalties in RL; consequently, policies may violate strict joint limits, balance requirements, or contact constraints, resulting in failure during physics-based execution. The temporal nature of the problem also introduces long-horizon credit assignment issues, where risk reduction depends on early preparatory actions (e.g., foot placement) that are difficult to reinforce due to accumulated drift over long subtasks. Finally, the sample and computational costs of physics-based long-horizon RL might be overwhelming, requiring extensive rollouts that make training and tuning prohibitively expensive.

To address these gaps, this study aims to explore the potential of RL for individually tailored task-posture optimization by developing and testing the feasibility of a pipeline that effectively mitigates the technical barriers.

Proposed Posture Optimization Technique

Figure 1 provides an overview of the proposed RL-based individually tailored and biomechanically feasible posture optimization pipeline.

Problem Formulation

Individualized task-posture correction is formulated as a long-horizon, safety-constrained, multi-objective RL problem executed in a physics simulator. Rather than synthesizing motion from scratch, the policy predicts bounded joint-space residuals relative to a demonstration trajectory. At each control step, the action is applied as a joint-wise offset to the current demonstration pose target (i.e., the reference pose used as the tracking setpoint) and is clipped using conservative per-joint limits, ensuring that the corrections remain local and physically trackable by the simulated body. The corrected target is executed in a contact- and stability-aware simulator, and the post-contact state is evaluated to generate a risk signal for reward shaping.

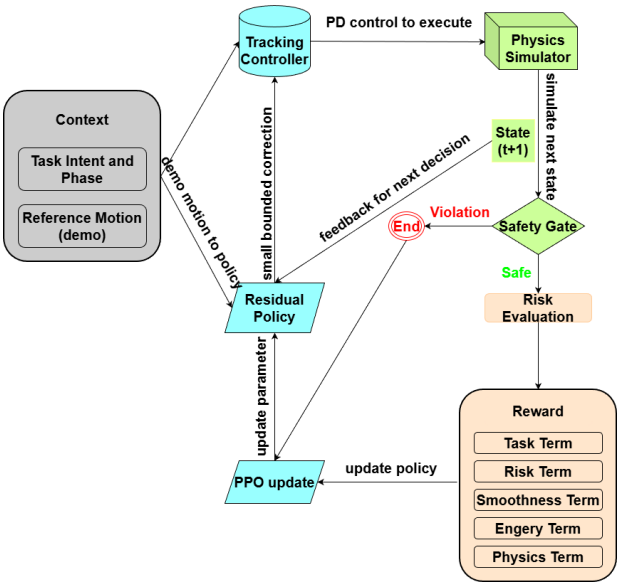

Figure 1: The proposed RL-based individually tailored and biomechanically feasible posture optimization pipeline.

To support distribution-aware risk shaping, the framework models both expected risk and the probability of high-risk events using a lightweight predictor distilled from a richer teacher model, making it practical for on-policy training. Safety is enforced through hard terminations triggered by simulator- and biomechanics-based checks (e.g., joint-limit violations, balance or contact failures, or spinal overload). Any such violation immediately terminates the episode and incurs a terminal penalty, thereby restricting exploration to physically executable behaviors. Finally, to discourage unrealistic reward shortcuts, policy updates are regularized toward a pretrained motion diffusion prior through a KL-divergence regularizer, encouraging outputs to remain within the manifold of human motion.

State Representation

A standardized kinematic encoding with explicit temporal context is used for state representation. All kinematic quantities are expressed in a consistent world frame and normalized using dataset-level statistics, while velocities are scaled by the control timestep to keep magnitudes comparable across terms. Each per-frame state s_t includes: (i) the root pose (position and orientation) and root linear/angular velocities; (ii) per-joint local rotations and angular velocities; and (iii) task-context features, including the goal pose, a phase ID (represented as a one-hot stage token), and a short reference snippet from the demonstration motion. The phase ID provides explicit stage information for indexing time-varying reference targets and disambiguating similar poses across different phases, which helps stabilize long-horizon imitation-style control (Peng

et al., 2018). Goal features are encoded as transforms relative to the pelvis (translation + 6D rotation). The reference snippet serves as a local motion anchor (e.g., a short horizon of target joint poses/velocities) to preserve task intent, while the policy is restricted to bounded corrections rather than free-form resynthesis.

Joint rotations are represented using the continuous 6D rotation encoding, defined by the first two columns of the rotation matrix, which avoids the singularities and discontinuities associated with Euler angles during learning. To provide short-term temporal context for contact and stability transitions and delayed biomechanical responses, the policy is conditioned on a stacked history window, where t indexes control timestep, H denotes the number of stacked frames, and s_k is the per-frame state at timestep k

$$\tilde{s}_t \triangleq [s_{t-H+1}, s_{t-H+2}, \ldots, s_t], \qquad H = 32. \tag{1}$$

With a 30 Hz control rate, $H = 32$ provides approximately 1.07 s of context while allowing the policy to remain feed-forward for stable on-policy learning. When personalization is enabled, the observation is augmented with subject descriptors (e.g., anthropometric measurements, range of motion indicators, and strength proxies), enabling capability-aware corrections rather than one-size-fits-all adjustments.

Action Representation

Actions are defined as continuous joint-space residuals (pose increments) bounded by joint-specific limits to enforce conservative, anatomically plausible corrections. Bounding is implemented through element-wise clipping (or tanh squashing) within predefined per-joint ranges. At each control step, the correction is added to the demonstration pose target and tracked by a stabilized PD controller in the physics simulator. Tracking is performed in joint space under torque limits and damping regularization to reduce overshoot and improve stability during contact switching. These bounds prevent implausible pose jumps and confine the corrections to a local feasible regime.

Given (s_t, a_t), where a_t is the pose correction applied at control timestep t, the simulator executes the correction-adjusted tracking command and advances the system:

$$s_{t+1} = F(s_t, a_t), \tag{2}$$

where $F(\cdot)$ denotes one physics rollout step, including contact resolution and internal integration substeps. After the transition, the risk student is queried using the updated history window $\tilde{s}_{t+1}$ (which includes the realized s_{t+1}). Evaluating risk after contact resolution and stabilization ties the reward signal to physically realized outcomes rather than to a pre-contact kinematic estimate and yields more consistent credit assignment when small corrections trigger compensation.

Risk-Awareness via Student–Teacher Distillation

The diffusion-based teacher builds upon the prior multimodal dynamic MSD-risk assessment framework, which fuses (i) skeleton dynamics, (ii) rule-based ergonomic indicators such as REBA-related cues and L5/S1-related loading proxies, (iii) visual-context features, and (iv) subject-related factors into a distributional risk representation. In the present pipeline, this upstream teacher provides supervisory targets for the mean risk and the probability of high-risk exceedance used in RL reward shaping. The upstream MSD risk evaluator represents uncertainty through a diffusion-based head and outputs a distribution over a scalar risk intensity u. Here, u denotes a normalized MSD-risk score derived from the upstream multimodal risk model. To make this risk signal numerically stable and comparable across motions and subjects during RL training, the raw teacher-side risk outputs are mapped to $[0, 1]$ using a monotonic calibration function fitted on the teacher training set. This mapping preserves the relative ordering of risk levels while placing them on a common scale for reward computation. Mean risk alone can obscure brief but hazardous spikes; therefore, the reward also penalizes the probability of exceeding a high-risk threshold. A tail event is defined using a fixed high-risk threshold $u_{\mathrm{hi}} = 0.8$, and the exceedance probability $\rho_t = \Pr(u \geq u_{\mathrm{hi}})$ is computed. This tail probability is explicitly penalized to discourage postures that retain non-negligible probability mass in the high-risk regime, thereby inducing risk-averse behavior. The threshold is selected to represent an upper-risk regime under the calibrated scale while keeping tail events sufficiently frequent for stable learning.

Under physics-based execution, contact events and safety terminations can abruptly end rollouts and amplify reward noise over long horizons. Proximal Policy Optimization (PPO) is adopted as the learning backbone because it provides stable, reliable optimization for long-horizon continuous control in such settings. However, PPO is on-policy and requires frequent rollout collection with step-wise reward evaluation, making the per-step cost of computing the risk signal a practical bottleneck. Estimating the tail probability via diffusion sampling at every environment step would be computationally prohibitive in the online PPO loop. This cost is removed from online training via teacher–student distillation: offline, a diffusion teacher produces distributional supervision targets, while online, a lightweight student predicts only the risk statistics required by the RL objective:

$$(\hat{\mu}_t, \hat{\rho}_t) = \mathrm{RiskStudent}(\tau_{t+1}). \tag{3}$$

This is sufficient because the RL objective depends on the expected risk level and the probability of entering the high-risk regime, rather than the full risk distribution. Here, $\hat{\mu}_t$ approximates the teacher mean risk and $\hat{\rho}_t$ approximates the teacher tail probability. The student is trained offline by MSE regression to teacher-derived targets, where the teacher's μ_t and ρ_t are estimated from diffusion samples per window (e.g., $S = 8$). During RL rollouts, diffusion sampling is not performed; the diffusion teacher is used only for evaluation and ablation studies.

Multi-Objective Reward & Safety Constraints

A unified multi-objective reward is optimized under hard safety enforcement:

$$r_t = \sum_{\star \in \Omega} w_\star(t) R_\star(t) - \lambda c_t, \quad \Omega = \{\text{task}, \text{risk}, \text{smooth}, \text{energy}, \text{physics}\}. \tag{4}$$

The weights follow a fixed, deterministic curriculum, and λ is a safety multiplier. The curriculum emphasizes task fidelity and smoothness early in training (to preserve intent and stability) and gradually shifts weight toward risk reduction once reliable tracking is achieved. The indicator $c_t \in \{0, 1\}$ denotes a safety violation; when $c_t = 1$, the episode terminates immediately with an explicit terminal penalty. In implementation, the safety penalty is applied as a terminal cost with magnitude $P_{\text{term}} = 100$ (i.e., $-\lambda c_t$ is incurred only at termination).

The risk objective penalizes both the expected risk and the high-risk exceedance probability:

$$R_{\text{risk}}(t) = -\hat{\mu}_t - \alpha_{\text{tail}} \hat{\rho}_t. \tag{5}$$

where $\hat{\mu}_t$ is the predicted mean risk and $\hat{\rho}_t$ is the tail probability under the calibrated risk distribution.

The remaining objectives promote task fidelity and motion quality. The task term penalizes deviations from task intent and reference execution (e.g., goal-reaching errors, phase-dependent pose constraints, and trajectory deviation from the demonstration). The smoothness term discourages abrupt velocity and acceleration changes to suppress jerk-like corrections. The energy term penalizes actuation effort (e.g., the squared torque norm). The physics-realism term penalizes simulator-flagged artifacts such as stance-foot slip and excessive foot skating. Each term is normalized to a consistent scale before weighting to keep objectives comparable.

Safety termination is triggered by joint-limit exceedance, L5/S1 overload (excessive compressive loading at the L5/S1 junction, commonly used as a proxy for low-back injury risk), or balance/contact failure. Balance/contact failure is detected using simulator-resolved events such as falls, loss of required support contacts, or the center of mass leaving the support polygon. The L5/S1 load used for termination is computed by the physics/biomechanics engine, while the learned risk model is used only for reward shaping.

Training Loop

Training alternates between physics rollouts and policy updates. Offline, a diffusion-based risk teacher is trained and distilled into a lightweight risk student, and a separate motion diffusion model is trained as a realism prior. Online, each PPO iteration collects N episodes in the physics simulator using the stacked history input τ_t. After each transition, the risk student is queried on the updated window τ_{t+1} (which includes the realized s_{t+1}) and outputs $(\hat{\mu}_{t+1}, \hat{\rho}_{t+1})$ for distribution-aware reward shaping; hard violations trigger immediate termination with the terminal penalty. For brevity, the post-transition estimates $\text{RiskStudent}(\tau_{t+1})$ are denoted as $(\hat{\mu}_{t+1}, \hat{\rho}_{t+1})$. The policy is then updated using PPO for continuous control, with an additional KL regularizer between the policy distribution $\pi_\theta(a_t \mid \tau_t)$ and the prior distribution $\pi_{\text{prior}}(a_t \mid \tau_t)$, encouraging corrections to remain close to natural motion while still enabling risk reduction. The value function is updated by regression. Throughout training, a fixed deterministic curriculum schedules the objective weights $w_\star(t)$ to improve reproducibility and prevent weight oscillations under noisy on-policy returns.

Tests of the Proposed Technique

The framework was evaluated in physics-based, construction-like scenes along three criteria: (i) MSD risk reduction, measured in a distribution-aware manner (both the average risk level and the frequency of high-risk spikes via exceedance probability), (ii) task and motion preservation, measured by task-fidelity errors and motion-quality statistics relative to the reference motion, and (iii) safety and feasibility, measured by violation rates and rollout success/failure.

Test Setup

All evaluations were performed on a desktop with an NVIDIA RTX 5090 (32 GB VRAM), 64 GB RAM, and an Intel Core Ultra 7 265K CPU. Policies were trained in PyTorch using PPO, with rollouts executed in Isaac Lab (NVIDIA Omniverse Isaac Sim / PhysX). The simulator used $\Delta t_{\text{phys}} = 1/120$ s with decimation $d = 4$, yielding a 30 Hz control rate ($\Delta t_{\text{ctrl}} = 1/30$ s) and an episode horizon of $H = 240$ control steps (8 s). The default Isaac Lab HumanoidAMP configuration ($n_q = 28$ actuated DOFs) was used with joint limits enabled; reference targets were tracked via stabilized PD control with torque limits, using the default HumanoidAMP gains fixed across methods for fair comparison.

Results were averaged over 5 training seeds. Each seed was evaluated under 40 randomized initializations (200 runs total); metrics were averaged within each seed and reported as mean ± standard deviation (SD) across seeds. PPO updates used 4096 environment steps per iteration collected from parallel rollouts; episodes could terminate early due to hard-safety constraints, so sampling continued until the step budget was reached. Standard continuous-control PPO hyperparameters were used ($\gamma = 0.99$, $\lambda_{\text{GAE}} = 0.95$, $\epsilon = 0.2$, learning rate 3×10^{-4}, 10 epochs per update, minibatch size 256) to avoid tuning to a specific scene configuration. The risk student was queried at each timestep on the post-transition window τ_{t+1} (including the realized s_{t+1}) for reward shaping, with negligible overhead relative to physics simulation.

Test Data

Reference motion clips were drawn from AMASS (Mahmood et al., 2019), a large unified corpus of MoCap sequences represented in a common model parameterization (e.g., SMPL), which provides large-scale MoCap sequences in a standardized body-model parameterization (e.g., SMPL) for consistent kinematic processing and retargeting. Action semantics and temporal segmentation were obtained from BABEL (Punnakkal et al., 2021), which supplies sequence- and frame-level action labels aligned with AMASS motions. Using BABEL labels, clips relevant to construction-like activities were selected (e.g., turning locomotion, bending/stooping, crouching, reaching, and overhead arm motion). Sequences were resampled to 30 Hz, segmented into 8 s clips (240 frames), and retargeted to the simulator humanoid via offline constrained inverse kinematics with joint-limit enforcement (joint mapping, scale normalization, and root alignment). Root motion was normalized (heading alignment and floor-height correction) to enable consistent reuse across scene layouts and randomized initializations.

Test Configuration

Episodes were initialized from reference clips and executed in parameterized construction-like layouts (e.g., corridor-to-corner navigation and low-clearance passage traversal) that require physically feasible locomotion and posture adaptation. Task objectives were specified by scene geometry and goal conditions, such as reaching a target region, progressing along a prescribed path, and satisfying clearance constraints. Scene parameters were randomized across episodes (e.g., corridor width, turn angle, and target location) to assess robustness under layout variation. To induce MSD-relevant mechanical demand without object handling, an equivalent virtual payload was applied during rollouts (e.g., a constant downward load at the hand frame), producing trunk moments and L5/S1 loading during stooping and reaching.

Representative examples of whole-body postures relevant to these task settings are shown in Figure 2. The three poses are included to visually clarify the representative posture demands and whole-body configurations considered in this study.

Baselines and Metrics

The technique is compared against three representative baselines: (i) reference tracking only (RTO), (ii) PPO without risk shaping (SPO), and (iii) mean-risk-only PPO (tail term removed). To examine whether subject-specific inputs contribute to individualized optimization, we additionally include a descriptor-ablated variant of the proposed method, in which subject descriptors are removed while all other components and training settings are kept unchanged. This comparison is used as a targeted individualization check rather than as a separate baseline family. Reported metrics include distributional risk (mean risk and exceedance probability), safety/feasibility (success rate), task fidelity (goal-reaching error and trajectory deviation), motion quality (smoothness and correction magnitude), and effort (mean squared torque, $\|\tau\|_2^2$). Results are reported for (1) overall effectiveness under physics-based execution (risk reduction with task fidelity and feasibility), (2) risk aversion by removing the tail term (change in high-risk exceedance), and (3) an individualization check based on the descriptor-ablated comparison.

Figure 2: Representative whole-body posture categories relevant to task-motion optimization.

Results and Discussions

As shown in Table 1, the proposed technique achieved the best overall trade-off under physics-based execution. It reduced mean risk from 10.88 (RTO) and 9.93 (SPO) to 6.15, and lowered tail probability from 0.25 to 0.14, while maintaining strong task completion (success 0.85) and the lowest task error (0.19). Corrections remained bounded (residual magnitude 0.19) and smoothness stayed comparable to the baselines (1.03 vs. 1.00–1.10), supporting local, physically plausible posture adaptation rather than rewriting the reference motion.

Table 1: Overall effectiveness

Metrics	RTO	SPO	PPO	Proposed
Mean risk	10.88 ± 0.27	9.93 ± 0.33	8.32 ± 0.26	**6.15 ± 0.23**
Tail prob.	0.25 ± 0.01	0.22 ± 0.02	0.20 ± 0.02	**0.14 ± 0.01**
Task error	0.38 ± 0.01	0.29 ± 0.01	0.23 ± 0.01	**0.19 ± 0.01**
Success	0.62 ± 0.02	0.82 ± 0.02	0.84 ± 0.02	**0.85 ± 0.01**
Res. mag.	0.00	0.15 ± 0.01	0.21 ± 0.02	**0.19 ± 0.01**
Smoothness	1.00 ± 0.03	1.07 ± 0.05	1.10 ± 0.04	**1.03 ± 0.03**

The reduction in risk (without compromising success rates) demonstrates that the hard safety constraints and physics-based simulation act as a critical filter, ensuring that all suggested posture corrections remain within the user's biomechanical limits. Unlike the SPO baseline, which fails to prioritize health, or the PPO baseline, which lacks robustness against extreme risk, the proposed method identifies a "safe manifold" for human motion.

Furthermore, the low residual magnitude (0.19) and high task fidelity (0.19 error) confirm that the technique achieves optimization by actively incorporating the user's habitual motion patterns. This suggests that the pipeline

does not impose a generic, "one-size-fits-all" posture; instead, it provides a personalized refinement that preserves the worker's unique motor profile while systematically hedging against both average and tail-end MSD risks. Consequently, the framework proves to be a viable bridge between complex biomechanical theory and practical, actionable interventions in labor-intensive environments.

Beyond improving average ergonomic scores, the necessity of this distribution-aware approach is most clearly evidenced by its performance in extreme scenarios. As shown in Table 2, the proposed technique significantly reduced rare-but-severe outcomes that standard RL methods often overlook. While the SPO and PPO baselines exhibited high-risk episode frequencies of 0.34 and 0.29, respectively, the proposed framework lowered this incidence to 0.12, flattening the tail of the risk distribution.

Table 2: Tail-sensitive outcomes for risk-aversion ablation

Method	P95 risk	High-risk episodes
SPO	12.18 ± 0.42	0.34 ± 0.05
PPO	11.64 ± 0.47	0.29 ± 0.04
Proposed	**9.63 ± 0.27**	**0.12 ± 0.02**

The reduction of P95 risk to 9.63 (compared to 12.18 for SPO and 11.64 for PPO) further confirms that the pipeline does not just lower the mean burden but proactively avoids the most hazardous postural configurations. This demonstrates that by penalizing tail-event probabilities through the distilled risk student, the system provides a level of safety reliability that is essential for real-world deployment, where injury prevention must be absolute. These findings ultimately validate that the proposed RL-based pipeline effectively hedges against the technical barriers of risk-reward instability and safety enforcement in complex task environments.

Table 3: Subject-conditioning ablation

Method	Mean risk	Tail prob.	Success
W/o Subject	9.92 ± 0.41	0.23 ± 0.02	0.81 ± 0.04
Proposed	**6.15 ± 0.23**	**0.14 ± 0.01**	**0.85 ± 0.01**

As shown in Table 3, the descriptor-ablated variant produced a substantially higher mean risk (9.92 ± 0.41) than the full subject-conditioned policy (6.15 ± 0.23), while task success remained broadly comparable (0.81 ± 0.04 vs. 0.85 ± 0.01). The relatively small change in success is likely because the evaluated task settings are not highly failure-prone, so the main effect of subject conditioning is expressed more clearly through risk reduction than through task completion rate. This result suggests that subject-specific descriptors contribute materially to the optimization outcome rather than acting as inactive auxiliary inputs. The pronounced increase in risk after removing individual information indicates that the learned correction strategy is not purely generic but is meaningfully shaped by subject-conditioned features. These findings provide additional evidence for the individualized character of the proposed optimization framework.

Conclusions

An RL-based pipeline is presented for individualized posture correction under physics-based execution and hard safety constraints. Rather than synthesizing motion from scratch, the policy outputs bounded joint-space residuals that produce local, actionable corrections while preserving task intent. A distilled risk student provides mean risk and high-risk exceedance probability for distribution-aware reward shaping, enabling explicit suppression of tail events beyond mean-only optimization.

The tests demonstrate that this design effectively reduces MSD risk while maintaining high task success and physical feasibility under the given task settings. The added subject-conditioning ablation further shows that removing individual descriptors substantially increases mean risk without yielding a meaningful improvement in task success, indicating that subject-specific inputs contribute materially to the learned correction strategy and supporting the individualized nature of the proposed framework. Ablations further indicate that hard safety enforcement and tail-risk shaping are key to stabilizing long-horizon rollouts under biomechanical constraints.

These findings indicate that integrating temporal context with distribution-aware reward shaping allows RL to generate corrections that are both ergonomically superior and physically executable. Ultimately, this research provides a scalable pathway for transitioning from generic ergonomic guidelines to personalized, biomechanically coherent digital interventions that can reduce the long-term burden of MSDs in labor-intensive industries. Future work will extend the framework to more contact-rich activities and integrate higher-fidelity musculoskeletal models to better support real-world ergonomic interpretation.

Acknowledgments

This work was supported by the New Frontiers in Research Fund (NFRF), grant number NFRFE-2023-00174. The authors wish to thank their industry partners for their help with data collection, as well as the anonymous subjects who participated in the data collection.

References

Belbeck, A., Cudlip, A. C., and Dickerson, C. R. (2014). Assessing the interplay between the shoulders and low back during manual patient handling techniques in a nursing setting. International Journal of Occupational Safety and Ergonomics, 20(1):127–137.

Chen, Y.-L. and Zhang, L.-P. (2025). Postural variability in sitting: Comparing comfortable, habitual, and correct strategies across chairs. Applied Sciences, 15(13):7239.

Ciccarelli, M., Germani, M., Marinelli, F., Papetti, A., and Pizzuti, A. (2026). An ergonomic zone polyhedral representation-based mathematical program to prevent work-related musculoskeletal risks. Expert Systems with Applications, 302:130579.

Dafarra, S., Romualdi, G., and Pucci, D. (2022). Dynamic complementarity conditions and whole-body trajectory optimization for humanoid robot locomotion. IEEE Transactions on Robotics, 38(6):3414–3433.

González-Alonso, J., Martín-Tapia, P., González-Ortega, D., Antón-Rodríguez, M., Díaz-Pernas, F. J., and Martínez-Zarzuela, M. (2025). Me-ward: A multimodal ergonomic analysis tool for musculoskeletal risk assessment from inertial and video data in working places. Expert Systems with Applications, 278:127212.

Hignett, S. and McAtamney, L. (2000). Rapid entire body assessment (reba). Applied ergonomics, 31(2):201–205.

Hulshof, C. T., Pega, F., Neupane, S., Colosio, C., Daams, J. G., Kc, P., Kuijer, P. P., Mandic-Rajcevic, S., Masci, F., van der Molen, H. F., et al. (2021). The effect of occupational exposure to ergonomic risk factors on osteoarthritis of hip or knee and selected other musculoskeletal diseases: A systematic review and meta-analysis from the who/ilo joint estimates of the work-related burden of disease and injury. Environment International, 150:106349.

Ji, X., Littman, A., Hettiarachchige, R. O., and Piovesan, D. (2023). The effect of key anthropometric and biomechanics variables affecting the lower back forces of healthcare workers. Sensors, 23(2):658.

Kee, D. (2022). Systematic comparison of owas, rula, and reba based on a literature review. International journal of environmental research and public health, 19(1):595.

Mahmood, N., Ghorbani, N., Troje, N. F., Pons-Moll, G., and Black, M. J. (2019). Amass: Archive of motion capture as surface shapes. In Proceedings of the IEEE/CVF international conference on computer vision, pages 5442–5451.

McAtamney, L. and Corlett, E. N. (1993). Rula: a survey method for the investigation of work-related upper limb disorders. Applied ergonomics, 24(2):91–99.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. nature, 518(7540):529–533.

Nadon, A. L., Cudlip, A. C., and Dickerson, C. R. (2018). Joint moment loading interplay between the shoulders and the low back during patient handling in nurses. Occupational Ergonomics, 13(1_suppl):81–90.

Peng, X. B., Abbeel, P., Levine, S., and Van de Panne, M. (2018). Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. ACM Transactions On Graphics (TOG), 37(4):1–14.

Punnakkal, A. R., Chandrasekaran, A., Athanasiou, N., Quiros-Ramirez, A., and Black, M. J. (2021). Babel: Bodies, action and behavior with english labels. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 722–731.

Ren, K. (2025). Feasibility of an integrated multi-model approach for dynamic musculoskeletal disorder risk assessment.

Rören, A., Ogiez, R., Gajny, L., Blasco, A., Bouvier, F. M., Feydy, A., Rannou, F., Lefèvre-Colau, M.-M., and Roby-Brami, A. (2024). Arm elevation involves changes in the whole spine: an exploratory study using eos imaging. BMC Musculoskeletal Disorders, 25(1):993.

Santos, W., Lorente, A., Rojas, C., Isidoro, R., Dias, A., Mariscal, G., Zabady, A. H., and Lorente, R. (2025). A systematic review and meta-analysis on the prevalence and demographic risk factors of work-related musculoskeletal disorders in construction workers. Frontiers in Public Health, 13:1651921.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. (2015). Trust region policy optimization. In International conference on machine learning, pages 1889–1897. PMLR.

Westermann, K., Lin, J. F.-S., and Kulić, D. (2020). Inverse optimal control with time-varying objectives: application to human jumping movement analysis. Scientific reports, 10(1):11174.

Zheng, Y. and Yumane, K. (2015). Human motion tracking control with strict contact force constraints for floating-base humanoid robots. US Patent 9,120,227.