REPO: REFERENCE-RELATIVE POLICY OPTIMIZATION FOR HIGH-FIDELITY HUMANOID MOTION IMITATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Physics-based humanoid controllers trained with Proximal Policy Optimization (PPO) have achieved high success rates in tracking diverse reference motions. However, successful execution does not necessarily imply faithful imitation: policies can increase cumulative rewards by surviving longer while adopting safer behaviors that deviate from the reference. To address this mismatch, we propose Reference-Relative Policy Optimization (RePO), which combines a critic-based viability advantage to support robust, long-horizon physical execution with a group-relative fidelity advantage to encourage faithful reference tracking. Instead of assigning a coarse trajectory-level score to each rollout, RePO normalizes returns-to-go across executions of the same motion at each aligned reference timestep, yielding dense, step-wise group-relative advantages that are blended with the viability advantage for policy updates. To reduce interference from survival differences, we introduce a survival gate that activates the fidelity advantage only when group-level execution is sufficiently reliable. We further adopt a kinematic cold start to initialize the policy from motion data, providing a reference-following prior for early group comparisons. Experiments on AMASS demonstrate improved motion fidelity and execution success without PHC's progressive training strategy. Furthermore, we demonstrate that our approach can facilitate the learning of complex motions. We validate this capability through challenging human–object interaction experiments on OMOMO, where RePO enables direct full-dataset training while the InterMimic fails to learn an effective policy.

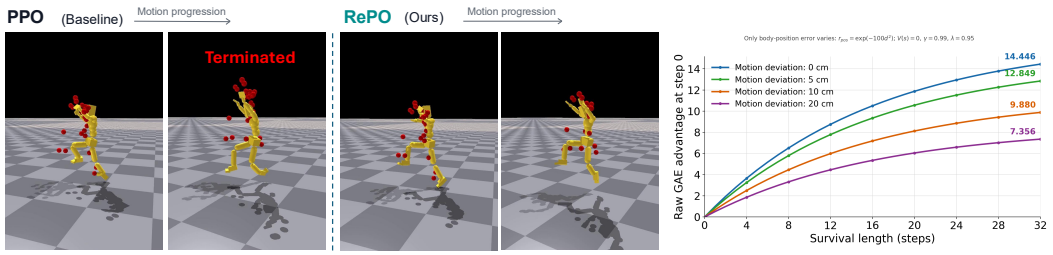
1 INTRODUCTION

Physics-based humanoid motion imitation aims to turn kinematic reference motions into physically executable behavior. Unlike purely kinematic reconstruction, it must account for body dynamics and contact constraints throughout execution. Proximal Policy Optimization (PPO) (Schulman et al., 2017) has become a widely used training framework for this task. Advances in motion priors and representations (Peng et al., 2022; Luo et al., 2023a; Xu et al., 2026a), reference-conditioned control and motion refinement (Luo et al., 2023b; Tessler et al., 2024; Xu et al., 2026b), and interaction learning (Xu et al., 2025; Lionar & Lee, 2026) have expanded the range of motions that humanoids can execute reliably.

However, a preference for stable execution can lead to conservative behaviors that sacrifice motion fidelity without ensuring task completion. As illustrated in Figure 1(a), PHC (Luo et al., 2023b) executes an insufficient jump rather than reproducing the reference takeoff dynamics. The resulting tracking error triggers early termination. This failure illustrates the pitfall of overly conservative behavior: favoring stability over the reference dynamics can ultimately undermine, rather than support, successful execution.

This discrepancy is closely tied to the optimization incentives induced by cumulative tracking rewards. With positive tracking rewards and early termination, failure removes all subsequent rewards, allowing longer survival to compensate for less accurate tracking. Figure 1(b) illustrates this effect through a simplified advantage calculation with a zero-valued critic: a less faithful rollout can receive a higher initial advantage than a more accurate one if it lasts longer. Consequently, a policy

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) (b)

Figure 1: Reward hacking and survival bias in PPO-based imitation. Red dots denote the target positions from the ground-truth motion. (a) The PPO baseline exhibits reward hacking: to preserve stability, it executes an insufficient jump rather than matching the target dynamics, incurring severe tracking error relative to the reference that ultimately triggers early termination. In contrast, RePO faithfully executes the full takeoff dynamics and completes the task. (b) Advantage estimates scale with survival length regardless of tracking accuracy, systematically rewarding survival over fidelity.

can improve its objective by adopting safer motion variants rather than reproducing the reference more faithfully. This raises the central question of our work: *How can policy optimization promote faithful imitation while sustaining physically viable execution?*

As a response to this question, we propose Reference-Relative Policy Optimization (RePO), which organizes policy improvement around a **viability advantage** and a **fidelity advantage**. The former draws on critic-based value estimates and future rewards (Schulman et al., 2015) to support stable, long-horizon execution. The latter introduces group-relative comparisons (Shao et al., 2024) among multiple executions of the same reference motion. Whereas the viability signal evaluates outcomes relative to critic-predicted state values, the fidelity signal evaluates them relative to alternative realizations of a shared target. We use a weighted combination of the two advantages in a common clipped policy objective, coordinating physical viability and motion fidelity through complementary learning signals.

Directly combining the two advantages is far from straightforward. First, conventional group-relative estimation assigns a single score to an entire rollout, obscuring variations in imitation quality across motion phases. As a result, well-executed and poorly tracked phases receive the same relative feedback, limiting phase-specific guidance for policy updates. Second, group-relative comparisons remain sensitive to differences in survival length: longer-lasting rollouts can accumulate higher returns despite less accurate tracking. The fidelity advantage may therefore reinforce survival-oriented updates rather than provide a complementary signal for faithful imitation.

To address the granularity mismatch, we introduce a **step-wise group-relative advantage**. We group rollouts of the same reference motion from a shared initial phase and normalize returns-to-go across group members at each aligned timestep, yielding dense, phase-specific relative feedback. To reduce interference from early termination, we introduce a Survival Gate (SG) that activates the fidelity advantage only when group-level execution is sufficiently reliable. We further introduce a kinematic cold-start strategy that pretrains the policy on digital human motion data, providing an initial motion prior for subsequent group-relative learning.

We evaluate RePO on AMASS, where it improves both motion fidelity and execution success without relying on PHC's progressive training strategy (Luo et al., 2023b) (see Figure 1(a), where RePO achieves full takeoff and completes the task). On AMASS-Train, RePO increases the success rate from 94.5% to 96.0% while reducing global and local tracking errors by 11.1% and 14.0%, respectively. On AMASS-Test, RePO achieves an 89.9% success rate, up from 87.0%, while also reducing tracking errors. Further analysis shows that our approach facilitates learning more complex motions. We extend the evaluation to challenging, contact-rich human-object interaction on OMOMO using the InterMimic (Xu et al., 2025) framework. Notably, RePO can be trained directly on the full OMOMO dataset and achieves a 22.7% motion-imitation success rate, even in a setting where our InterMimic baseline fails to converge. This result highlights RePO's potential to reduce reliance on the common practice of training subset-specific policies and distilling them into a unified policy.

2 Related Work

2.1 PPO-based Physics-based Humanoid Control

Human motion data (Mahmood et al., 2019; Kulkarni et al., 2024) is typically kinematic and does not directly match with the physically humanoid agents, due to differences in morphology, joint degrees of freedom, and motion dynamics. Physics-based reinforcement learning therefore provides a natural paradigm for motion imitation, where policies are optimized to track reference motions while respecting physical constraints. PPO (Schulman et al., 2017) has become the predominant optimization method for this setting. Building on this paradigm, prior work (Peng et al., 2022) has improved humanoid control through learned motion priors, latent motor representations (Xu et al., 2026a), physics-aware motion processing (He et al., 2024b;a), policy distillation (Xu et al., 2025), and task-specific reward design (Lionar & Lee, 2026; Xie et al., 2025), substantially enhancing robustness and motion coverage. These advances have further extended humanoid control from single-agent motion tracking to contact-rich interaction (**?**Li et al., 2026; Yu et al., 2026), multi-agent behaviors (Yao et al., 2026; Shibata et al., 2026), and highly dynamic whole-body skills (Xie et al., 2025; Han et al., 2025; Nam et al., 2026).

Despite these developments, the underlying policy optimization mechanism remains largely unchanged, with most methods still relying on the conventional PPO actor-critic formulation and critic-based advantage estimation. We revisit this criterion by introducing reference relative credit assignment, allowing policy improvement to be determined from alternative realizations of the same target motion.

2.2 Reinforcement Learning Policy Optimization

Reinforcement learning has become a powerful tool for optimizing task-specific policies and aligning models with human preferences. Proximal Policy Optimization (PPO) (Schulman et al., 2017) stabilizes policy updates through clipped importance ratios and typically combines a learned critic with Generalized Advantage Estimation (GAE) (Schulman et al., 2015). Group Relative Policy Optimization (GRPO)(Shao et al., 2024) instead estimates advantages through group-relative comparisons, avoiding an additional critic model. Recent works further improve policy optimization from different perspectives. Dr.GRPO (Liu et al., 2025) mitigates length bias in advantage estimation, DAPO (Yu et al., 2025) introduces asymmetric clipping and dynamic sampling, GSPO (Zheng et al., 2025) adopts sequence-level importance ratios to reduce variance, and SAPO (Gao et al., 2025) replaces hard clipping with smooth adaptive weighting for more stable and sample-efficient training. However, these methods are mainly designed for language-model post-training, with limited consideration of physics-based humanoid control. We instead introduce survival-gated, step-wise group-relative advantages for humanoid imitation, complementing critic-based credit assignment to improve motion fidelity.

3 Methods

This section presents Reference-Relative Policy Optimization (RePO). The overview of our approach is shown in Figure 2. We first review the PPO objective and critic-based advantage estimation in Section 3.1. Section 3.2 then introduces the hybrid formulation that combines viability and fidelity advantages. Within this formulation, we develop a step-wise group-relative estimator for phase-aligned comparisons, a survival gate for reliability-dependent activation, and a kinematic cold start for establishing an initial motion prior. We further describe KL regularization with an EMA-updated reference policy and conclude with the final actor objective.

3.1 Preliminary

Given rollouts collected by the behavior policy $\pi_{\theta_{\text{old}}}$, Proximal Policy Optimization (PPO) (Schulman et al., 2017) maximizes the clipped surrogate objective:

$$\mathcal{L}_{\text{PPO}}(\theta) = \mathbb{E}_t \left[\min \left(\rho_t(\theta) A_t^{\text{GAE}}, \text{clip} \left(\rho_t(\theta), 1 - \epsilon_{\text{clip}}, 1 + \epsilon_{\text{clip}} \right) A_t^{\text{GAE}} \right) \right], \tag{1}$$

where $\rho_t(\theta) = \pi_\theta(a_t \mid s_t) / \pi_{\theta_{\text{old}}}(a_t \mid s_t)$ is the policy probability ratio and ϵ_{clip} is the clipping parameter. The advantage is computed using Generalized Advantage Estimation (GAE) (Schulman

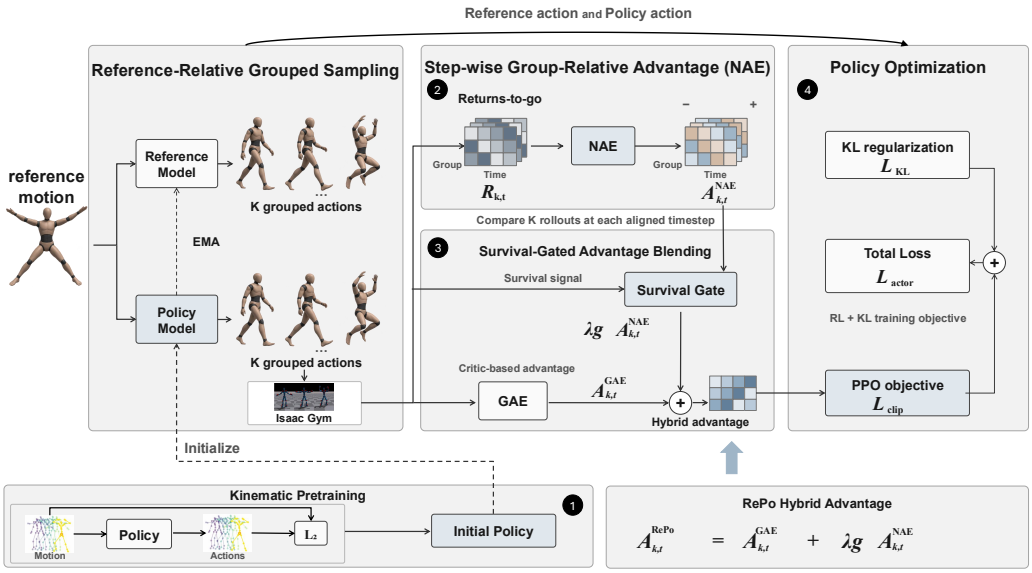


Figure 2: Overview of our approach. ❶ Kinematic pretraining initializes the policy for physics-based imitation. ❷ Grouped rollouts following the same reference motion enable step-wise group-relative advantage estimation through temporally aligned return-to-go comparisons. ❸ A survival gate regulates the contribution of NAE to the hybrid advantage. ❹ The training objective combines the PPO surrogate loss with KL regularization.

et al., 2015). With critic V_ψ, we define the temporal-difference residual as

$$\delta_t = r_t + \gamma V_\psi(s_{t+1}) - V_\psi(s_t), \tag{2}$$

and estimate the advantage as

$$A_t^{\mathrm{GAE}} = \sum_{\ell=0}^{T-t-1} (\gamma\lambda_{\mathrm{GAE}})^\ell \delta_{t+\ell}, \tag{3}$$

where r_t is the step reward, γ is the discount factor, λ_{GAE} controls the bias-variance trade-off, and T is the rollout length. The continuation value is zero at true terminal states. With positive tracking rewards, early termination removes future reward opportunities, allowing longer survival to compensate for poorer tracking. PPO can therefore favor physically stable but less faithful executions.

3.2 Reference-Relative Policy Optimization (RePO)

We propose RePO, a policy optimization framework that coordinates physical viability and motion fidelity in physics-based humanoid imitation. We combine critic-based GAE as the **viability advantage** to promote physically viable, long-horizon execution, with group-relative Normalized Advantage Estimation (NAE) (Shao et al., 2024) as the **fidelity advantage** to encourage faithful reference tracking.

For each reference motion, we sample a group of K rollouts, $\mathcal{G} = \{\tau_1, \ldots, \tau_K\}$, under the behavior policy $\pi_{\theta_{\mathrm{old}}}$. Let R_k denote the return of the k-th rollout, its group-relative advantage can be computed as

$$A_k^{\mathrm{NAE}} = \frac{R_k - \mu_R}{\sigma_R + \epsilon}, \tag{4}$$

where $\mu_R = \frac{1}{K}\sum_{j=1}^{K} R_j$ and $\sigma_R = \sqrt{\frac{1}{K}\sum_{j=1}^{K}(R_j - \mu_R)^2}$ are the within-group mean and standard deviation, and ϵ ensures numerical stability. We combine both signals into a hybrid advantage:

$$A_{k,t}^{\mathrm{RePO}} = A_{k,t}^{\mathrm{GAE}} + \lambda A_k^{\mathrm{NAE}}, \tag{5}$$

where λ controls the contribution of the group-relative signal. A_k^{NAE} is shared across all timesteps of rollout k. We next refine this trajectory-level estimate to provide step-wise relative supervision.

Step-wise group-relative advantage. To obtain phase-specific relative feedback, we replace the trajectory-level advantage A_k^{NAE} (Shao et al., 2024) with a step-wise estimate $A_{k,t}^{\mathrm{NAE}}$. We initialize all rollouts in a group from the same reference motion and initial phase, so aligned timesteps correspond to the same target. For rollout k, its discounted return-to-go at timestep t is defined as

$$R_{k,t} = \sum_{i=t}^{T-1} \gamma^{i-t} r_{k,i}, \tag{6}$$

where T denotes the rollout length. Using return-to-go allows the relative comparison to account for subsequent execution, rather than evaluating each step solely by its immediate reward. We then normalize these returns across group members at each timestep:

$$A_{k,t}^{\mathrm{NAE}} = \frac{R_{k,t} - \mu_t^R}{\sigma_t^R + \epsilon}, \tag{7}$$

where the timestep-specific group statistics are

$$\mu_t^R = \frac{1}{K}\sum_{j=1}^{K} R_{j,t}, \qquad \sigma_t^R = \sqrt{\frac{1}{K}\sum_{j=1}^{K}(R_{j,t} - \mu_t^R)^2}. \tag{8}$$

Each reference timestep thus has its own group-relative baseline, allowing a rollout's relative advantage to vary across motion phases instead of assigning the same feedback to all its actions. We substitute $A_{k,t}^{\mathrm{NAE}}$ for A_k^{NAE} in the hybrid advantage.

Survival-gated group-relative optimization. Even with temporally aligned comparisons, the fidelity advantage remains sensitive to survival length. Longer rollouts can accumulate higher returns despite poorer tracking, while early termination can reduce returns even when tracking before failure is accurate. Consequently, NAE may reinforce the survival preference already captured by GAE rather than provide a distinct signal for motion fidelity. Fidelity-oriented updates may also conflict with viability learning when executions remain dynamically unstable. We therefore introduce a survival gate to regulate the contribution of NAE according to group-level execution reliability. For rollout k, we define its normalized survival ratio as $q_k = \frac{L_k}{H_k}$, where L_k denotes the number of successfully executed timesteps and H_k is the corresponding reference-motion horizon. This normalization measures execution reliability relative to the target horizon, allowing a common criterion across motions of different lengths. We aggregate the rollout-level ratios into the group survival score $\bar{q} = \frac{1}{K}\sum_{j=1}^{K} q_j$ and define the gate as

$$g = \mathbb{I}(\bar{q} \geq \tau), \tag{9}$$

where τ is a reliability threshold (set to 0.9). We apply this gate across all group members and update Eq. (5) with the gated step-wise NAE:

$$A_{k,t}^{\mathrm{RePO}} = A_{k,t}^{\mathrm{GAE}} + \lambda g A_{k,t}^{\mathrm{NAE}}. \tag{10}$$

For groups below the threshold, the hybrid advantage reduces to GAE, prioritizing physically viable execution. For groups meeting the threshold, we add the step-wise NAE term to refine relative tracking quality while maintaining the viability signal. Because the gate is determined separately for each rollout group, fidelity refinement depends on the group's execution reliability rather than a fixed training-stage transition.

Kinematic cold start. Informative group-relative learning relies on meaningful variations in execution quality within each rollout group. With poorly initialized policies, rollouts may fail similarly or deviate substantially from the reference, leaving raw relative scores weakly informative about motion fidelity or dominated by incidental differences in survival. We therefore adopt a kinematic cold start to establish an initial reference-following prior. Before physics-based optimization, we pretrain the policy on purely kinematic motion data to learn a coarse mapping from reference-motion observations to target actions, without simulator interaction. We use the resulting parameters θ_{kin} to initialize the policy, providing a motion prior for subsequent group-relative learning. To regularize subsequent policy updates, we introduce a reference policy $\pi_{\theta_{\mathrm{ref}}}$, initialized as $\theta_{\mathrm{ref}}^{(0)} = \theta_{\mathrm{kin}}$. We penalize deviation from this reference whenever the fidelity advantage is activated:

$$\mathcal{L}_{\mathrm{KL}}(\theta) = \mathbb{E}_t\left[g D_{\mathrm{KL}}\left(\pi_\theta(\cdot \mid s_t) \| \pi_{\theta_{\mathrm{ref}}}(\cdot \mid s_t)\right)\right]. \tag{11}$$

Here, g is the survival gate for the corresponding rollout group, shared across its time steps. Using the same gate couples reference-policy regularization with fidelity-oriented updates, rather than imposing the kinematic prior unconditionally throughout training. A fixed kinematic reference can overly restrict adaptation to physical dynamics. We therefore update the reference parameters through an exponential moving average of the policy learned in simulation:

$$\theta_{\text{ref}}^{(n)} = \beta\theta_{\text{ref}}^{(n-1)} + (1-\beta)\theta^{(n)}, \tag{12}$$

where $\beta \in [0, 1)$ controls the averaging rate and $\theta^{(n)}$ denotes the current policy parameters at reference update n. The reference thus adapts to the evolving physics-based policy instead of permanently constraining optimization to the kinematic prior. Finally, we minimize the actor loss:

$$\mathcal{L}_{\text{actor}}(\theta) = -\mathcal{L}_{\text{clip}}(\theta; A^{\text{RePO}}) + \alpha\mathcal{L}_{\text{KL}}(\theta), \tag{13}$$

where $\mathcal{L}_{\text{clip}}(\theta; A^{\text{RePO}})$ denotes the clipped PPO objective from Eq. 1 evaluated with the gated hybrid advantage, and $\alpha \geq 0$ controls the strength of reference-policy regularization.

4 Experiments

We first describe the datasets and evaluation metrics in Section 4.1. Section 4.2 presents quantitative and qualitative comparisons on AMASS, followed by a complementary evaluation on contact-rich human–object interaction in OMOMO. We then examine learning performance on challenging motions in Section 4.3 and analyze the relationship between GAE and NAE during training in Section 4.4. Finally, Section 4.5 evaluates the contributions of individual components and the effects of group size and advantage-mixing weight.

4.1 Datasets and Evaluation Metrics

Datasets. We use AMASS to evaluate the effectiveness of RePO in improving motion fidelity and execution success. AMASS unifies diverse human motion-capture sequences under a common SMPL-based representation. Following the data-processing protocol of PHC (Luo et al., 2023b), we use 11,313 training sequences and 138 held-out test sequences.

As a complementary evaluation, we use OMOMO to assess RePO's applicability to more complex, contact-rich human-object interaction tasks. OMOMO contains approximately 10 hours of full-body manipulation motions involving 15 objects. We follow the evaluation protocol of InterMimic (Xu et al., 2025), where successful execution requires coordinated humanoid motion and object manipulation under physical constraints.

Evaluation metrics. We report the success rate (Succ) on both datasets, defined as the percentage of reference sequences completed without early termination under the corresponding evaluation protocol. For AMASS, we additionally report global MPJPE (E_g) and local MPJPE (E_l), which measure mean per-joint position errors in global and root-relative coordinates, respectively. Acceleration Distance (E_{acc}) and Velocity Distance (E_{vel}) measure discrepancies in joint accelerations and velocities between simulated and reference motions. For OMOMO, we report Duration, measuring how long reference tracking is maintained before failure or sequence completion, together with Human Tracking Error (E_h) and Object Tracking Error (E_o), which quantify deviations of the simulated humanoid and object trajectories from their references. Higher success rates and longer tracking durations, together with lower tracking errors, indicate better performance. For rows marked with an asterisk, tracking errors are computed only over successful executions.

4.2 Quantitative and Qualitative Comparisons

Table 1 compares PPO, GRPO, and RePO under the same PHC training framework. RePO achieves the best overall performance in both success rate and motion fidelity. GRPO yields a much lower success rate and larger overall errors. However, when considering only successful executions (marked with $*$), $\text{PHC}_{+\text{GRPO}}$ generally achieves the lowest tracking errors, indicating that GRPO strongly favors motion similarity but may sacrifice physical viability. Figure 3 qualitatively compares RePO with the PPO baseline. It can be seen that our method better tracks the ground-truth

Table 1: Quantitative comparison on the AMASS dataset. * indicates that errors are computed only over successfully tracked sequences.

Method	AMASS-Train					AMASS-Test				
	Succ ↑	E_g ↓	E_l ↓	E_{acc} ↓	E_{vel} ↓	Succ ↑	E_g ↓	E_l ↓	E_{acc} ↓	E_{vel} ↓
PHC	94.5%	38.8	30.1	3.8	5.1	87.0%	55.3	36.6	8.5	10.8
PHC$_{+\text{GRPO}}$	72.3%	43.3	33.2	5.0	6.3	23.2%	86.0	54.2	14.5	17.5
PHC$_{+\text{RePO}}$	**96.0%**	**34.5**	**25.9**	**3.2**	**4.6**	**89.9%**	**51.2**	**34.3**	**6.7**	**9.4**
PHC*	94.5%	35.5	27.9	3.2	4.6	87.0%	49.7	33.3	6.9	9.4
PHC$_{+\text{GRPO}}$*	72.3%	31.2	27.1	2.7	4.0	23.2%	38.3	29.2	4.4	6.6
PHC$_{+\text{RePO}}$*	96.0%	31.9	24.2	2.8	4.1	89.9%	46.3	30.9	5.5	8.2

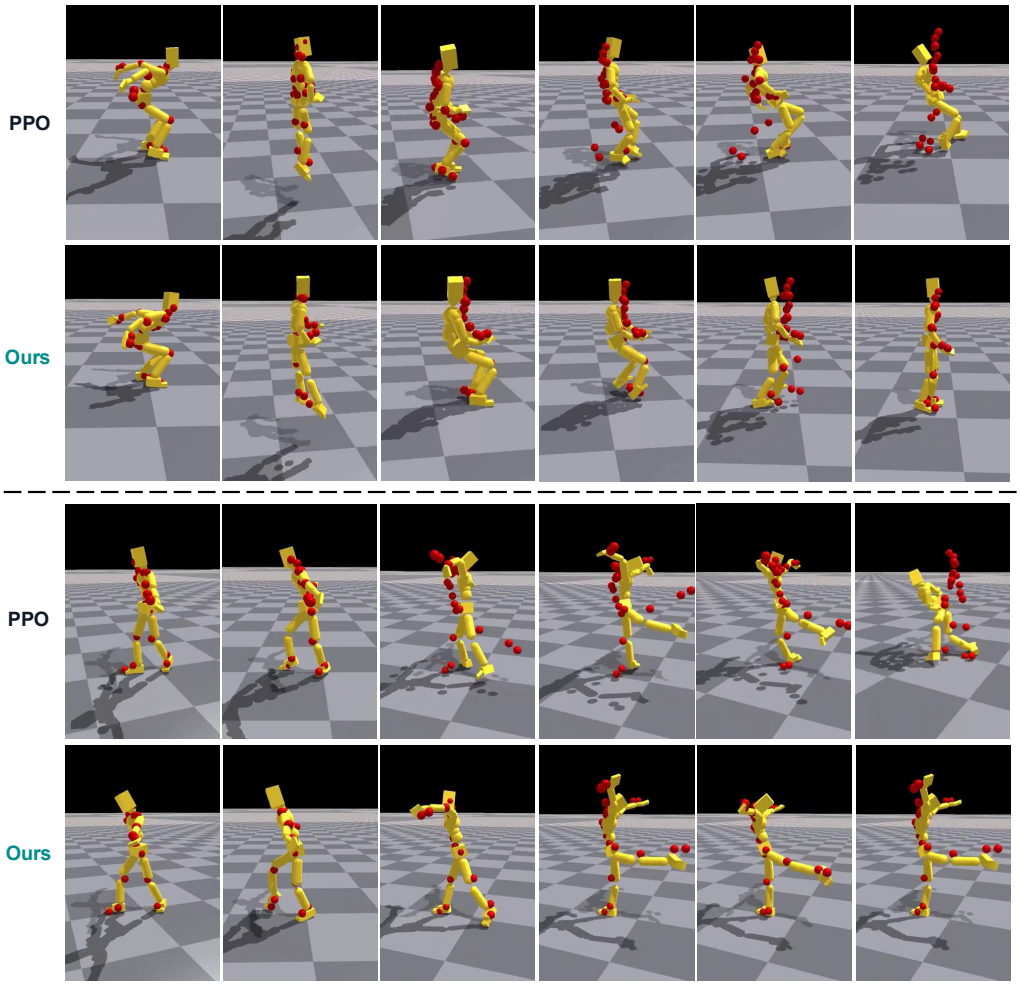


Figure 3: Visual results on the AMASS dataset. We compare our approach with PPO. Our method achieves successful execution while more closely matching the ground-truth motion. Note that more visual results are provided in the supplementary material.

motion (indicated by the red dots). In contrast, the PPO baseline (*i.e.*, PHC Luo et al. (2023b)) deviates substantially from the GT and ultimately fails (as shown in the final frame), further demonstrating that RePO effectively reconciles strict kinematic fidelity with robust physical execution.

We also evaluate on contact-rich human-object interaction using OMOMO. As shown in Table 2, InterMimic completely fails to converge under vanilla PPO (0.0% success rate). In contrast, RePO achieves a 22.7% success rate ($E_h = 0.113$, $E_o = 0.172$), demonstrating that our method effectively scales to complex contact dynamics.

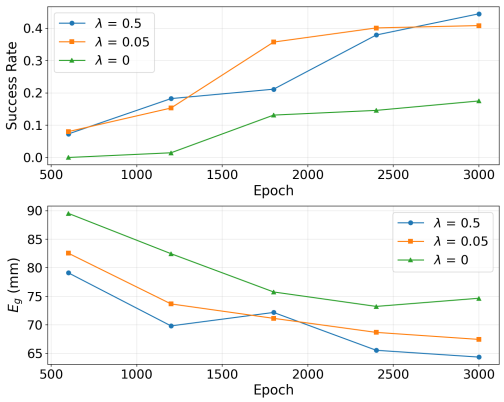


Figure 4: Impact of the advantage-mixing weight λ on tracking error (E_g) and success rate across challenging motion sequences.

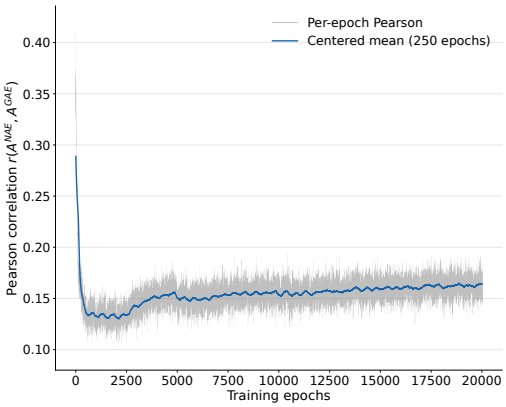


Figure 5: The Pearson Correlation between GAE and NAE Advantages during training.

4.3 Impact of Motion Fidelity on Physics-Based Motion Tracking

We collect 136 challenging motions from the AMASS-Train set that fail under both the vanilla PPO and our RePO methods. We control the strength of the fidelity-oriented optimization by adjusting the advantage-mixing weight λ, where a larger λ places more emphasis on tracking the reference motion, and $\lambda = 0$ collapses to standard PPO.

Table 2: Quantitative comparison on OMOMO. InterMimic fails to converge when trained directly on the full OMOMO dataset with PPO alone, whereas RePO partially overcomes this limitation.

Method	Succ ↑	Time ↑	E_h ↓	E_o ↓
Intermimic	0.0	NaN	NaN	NaN
Our	22.7%	93.29	0.113	0.172

As shown in Figure 4, increasing λ from 0 to 0.5 leads to clear changes in both tracking fidelity and imitation success. At 3,000 epochs, the global tracking error E_g decreases from 74.7 mm at $\lambda = 0$ to 67.5 mm at $\lambda = 0.05$ and 64.4 mm at $\lambda = 0.5$. Meanwhile, the corresponding success rate increases from 17.5% to 40.9% and 44.5%, respectively. Notably, even a small weight of $\lambda = 0.05$ reduces the tracking error by about 7 mm and more than doubles the success rate compared with vanilla PPO. These results suggest that, for challenging motions, a modest improvement in tracking fidelity can lead to a clear gain in success rate.

4.4 Understanding the Optimization Behavior of NAE

In this section, we analyze the relationship between GAE and NAE during training. To isolate the natural behavior of NAE from the survival gating mechanism, we measure the raw NAE before gating and compute its Pearson correlation with GAE. As shown in Figure 5, the correlation remains consistently low during training, indicating that NAE does not simply reproduce or rescale the optimization preference induced by GAE. Instead, the two estimators assign substantially different learning signals for the training samples.

Interestingly, the correlation is relatively high at the beginning of training but decreases rapidly thereafter. Early in training, large differences in survival length dominate both GAE and NAE, yielding similar optimization preferences. As the policy becomes more stable, survival provides less discriminative information, while NAE increasingly emphasizes relative tracking quality, resulting in a stronger divergence from GAE. In addition, we observe a moderate increase in correlation again during the later stage of training. One possible explanation is that improved motion fidelity gradually becomes associated with more stable physical execution as the policy converges, which is similar to our observation in Section 4.3. Nevertheless, the correlation remains limited, reflecting the inherent discrepancy between kinematic reference motions and their physically realizable counterparts.

4.5 Ablation Study

Ablation on the components. We ablate the key components of RePO, with results reported in Table 5. Removing the kinematic cold start reduces the success rate from 96.0% to 94.0% and

Table 3: The ablation study of the group size.

Group (G)	Succ ↑	E_g ↓	E_l ↓	E_{acc} ↓	E_{vel} ↓
2	94.9%	35.3	27.1	3.3	4.7
4	**96.0%**	**34.5**	**25.9**	**3.2**	**4.6**
8	94.0%	35.0	27.2	3.3	4.7
16	93.5%	35.6	27.8	3.6	5.0

Table 4: The ablation study of λ setting.

λ	Succ ↑	E_g ↓	E_l ↓	E_{acc} ↓	E_{vel} ↓
0.1	95.1%	36.2	27.9	3.7	5.0
0.5	**96.0%**	**34.5**	**25.9**	**3.2**	**4.6**
1.0	93.6%	35.0	26.9	3.3	4.6
2.0	90.6%	37.9	28.9	3.5	5.0

Table 5: The ablation study of components of RePO.

Methods	Succ ↑	E_g ↓	E_l ↓	E_{acc} ↓	E_{vel} ↓
PHC_{RePO}	**96.0%**	34.5	**25.9**	**3.2**	**4.6**
(−) Cold Start	94.0%	36.5	27.8	3.5	5.0
(−) Survival-Gated	94.6%	**34.3**	26.2	3.3	4.6
(−) Dense Advantage	91.6%	37.6	27.8	3.8	5.1

increases E_g and E_l from 34.5 to 36.5 and from 25.9 to 27.8, respectively. Without survival gating, the success rate drops to 94.6%. This suggests that suppressing the survival-length bias in GRPO advantages leads to better policy performance. Removing dense step-wise advantage estimation causes the largest degradation: the success rate falls to 91.6%, while E_g and E_{vel} increase to 37.6 and 5.1, respectively. These results highlight the importance of step-wise group-relative advantage, which provides distinct optimization signals for different phases of the reference motion.

Ablation on group size. We ablate the group size used for group-relative advantage estimation while keeping the total number of environments fixed for a fair comparison. As shown in Table 3, a small group size (e.g., $K = 2$) produces noisier group-relative statistics and less reliable advantage estimates. Compared with $K = 4$, $K = 2$ decreases the success rate by 1.1 percentage points and increases E_g and E_l by 0.8 and 1.2, respectively. In contrast, overly large groups (e.g., $K = 8$ or 16) reduce the diversity of reference motions within each training batch, leading to degraded performance, with the success rate decreasing from 96.0% to 94.0% and 93.5%, respectively. These results suggest that an appropriate group size should balance the reliability of intra-group relative estimation against the diversity of reference motions within each batch.

Ablation on blending ratio λ. We further ablate the NAE weight λ, with results reported in Table 4. Increasing λ from 0.1 to 0.5 improves the success rate from 95.1% to 96.0%, while reducing E_g and E_l from 36.2 to 34.5 and from 27.9 to 25.9, respectively. Further increasing λ to 1.0 and 2.0 lowers the success rate to 93.6% and 90.6%, respectively, with E_g and E_l worsening to 37.9 and 28.9 at $\lambda = 2.0$. This degradation suggests that excessive emphasis on alignment with kinematic references may conflict with physical viability and induce early termination. A moderate weight of $\lambda = 0.5$ achieves the best balance between motion fidelity and robust physical execution.

5 Conclusion

We presented RePO, a reference-relative policy optimization framework for faithful and robust physics-based humanoid motion imitation. By combining a critic-based viability advantage with a step-wise group-relative fidelity advantage, we coordinate physical execution and reference tracking. Survival gating and kinematic cold-start initialization support reliable relative comparisons during training. Experiments on AMASS demonstrate improvements in both motion fidelity and execution success without relying on PHC's progressive training strategy. We further demonstrate that RePO facilitates learning more complex motions through challenging human–object interaction on OMOMO, including direct full-dataset training where the InterMimic baseline fails to learn an effective policy. These results highlight the role of policy optimization in improving imitation quality and extending the range of learnable motions, while suggesting a path toward reduced reliance on subset-wise policy training and distillation.

AI USE STATEMENT

In this work, we used OpenAI's ChatGPT to assist with manuscript drafting and language refinement, including grammar, clarity, and readability. The authors reviewed and revised the AI-assisted text and take full responsibility for the final content, analyses, results, and conclusions.

REFERENCES

Chang Gao, Chujie Zheng, Xiongh ui Chen, Kai Dang, Shixuan Liu, Bowen Yu, An Yang, Shuai Bai, Jingren Zhou, and Junyang Lin. Soft adaptive policy optimization. *CoRR*, abs/2511.20347, 2025.

Jinrui Han, Weiji Xie, Jiakun Zheng, Jiyuan Shi, Weinan Zhang, Ting Xiao, and Chenjia Bai. Kungfubot2: Learning versatile motion skills for humanoid whole-body control. *CoRR*, abs/2509.16638, 2025.

Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris M. Kitani, Changliu Liu, and Guanya Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. In *8th Annual Conference on Robot Learning*, 2024a.

Tairan He, Zhengyi Luo, Wenli Xiao, Chong Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Learning human-to-humanoid real-time whole-body teleoperation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2024, Abu Dhabi, United Arab Emirates, October 14-18, 2024*. IEEE, 2024b.

Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, and Leonidas J. Guibas. NIFTY: neural object interaction fields for guided human motion synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 947–957. IEEE, 2024.

Xinyao Li, Xialin He, Runpei Dong, and Saurabh Gupta. Contactmimic: Humanoid object interaction via contact control. *CoRR*, abs/2607.08742, 2026.

Stefan Lionar and Gim Hee Lee. Teamhoi: Learning a unified policy for cooperative human-object interactions with any team size. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2026, Denver, CO, USA, June 3-7, 2026*, 2026.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *CoRR*, abs/2503.20783, 2025.

Zhengyi Luo, Jinkun Cao, Josh Merel, Alexander W. Winkler, Jing Huang, Kris Kitani, and Weipeng Xu. Universal humanoid motion representations for physics-based control. *ArXiv*, 2023a.

Zhengyi Luo, Jinkun Cao, Alexander W. Winkler, Kris Kitani, and Weipeng Xu. Perpetual humanoid control for real-time simulated avatars. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023b.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: archive of motion capture as surface shapes. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 5441–5450. IEEE, 2019.

Sanghyeok Nam, Byoungjun Kim, Daehyung Park, and Tae-Kyun Kim. Dynamic full-body motion agent with object interaction via blending pre-trained modular controllers. *CoRR*, abs/2605.11369, 2026.

Xue Bin Peng, Yunrong Guo, Lina Halper, Sergey Levine, and Sanja Fidler. ASE: large-scale reusable adversarial skill embeddings for physically simulated characters. *ACM Trans. Graph.*, 2022.

John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and P. Abbeel. High-dimensional continuous control using generalized advantage estimation. *CoRR*, abs/1506.02438, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *ArXiv*, abs/1707.06347, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Jun-Mei Song, Mingchuan Zhang, Y. K. Li, Yu Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *ArXiv*, 2024.

Yuto Shibata, Kashu Yamazaki, Lalit Jayanti, Yoshimitsu Aoki, Mariko Isogawa, and Katerina Fragkiadaki. Learning to assist: Physics-grounded human-human control via multi-agent reinforcement learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2026, Denver, CO, USA, June 3-7, 2026*, 2026.

Chen Tessler, Yunrong Guo, Ofir Nabati, Gal Chechik, and Xue Bin Peng. Maskedmimic: Unified physics-based character control through masked motion inpainting. *ACM Transactions on Graphics (TOG)*, 2024.

Weiji Xie, Jinrui Han, Jiakun Zheng, Huanyu Li, Xinzhe Liu, Jiyuan Shi, Weinan Zhang, Chenjia Bai, and Xuelong Li. Kungfubot: Physics-based humanoid whole-body control for learning highly-dynamic skills. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025.

Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, and Liangyan Gui. Intermimic: Towards universal whole-body control for physics-based human-object interactions. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Sirui Xu, Samuel Schulter, Morteza Ziyadi, Xialin He, Xiaohan Fei, Yu-Xiong Wang, and Liangyan Gui. Interprior: Scaling generative control for physics-based human-object interactions. *CoRR*, abs/2602.06035, 2026a. doi: 10.48550/ARXIV.2602.06035.

Weisheng Xu, Qiwei Wu, Jiaxi Zhang, Tan Jing, Yangfan Li, Yuetong Fang, Jiaqi Xiong, Kai Wu, Rong Ou, and Renjing Xu. Iterative closed-loop motion synthesis for scaling the capabilities of humanoid control. *CoRR*, abs/2602.21599, 2026b.

Wei Yao, Haohang Ma, Hongwen Zhang, Yunlian Sun, Liangjun Xing, Zhile Yang, Yuanjun Guo, Yebin Liu, and Jinhui Tang. Synagent: Generalizable cooperative humanoid manipulation via solo-to-cooperative agent synergy. *ArXiv*, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: an open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025.

Wongyun Yu, Young-Won Kim, and Minsu Cho. Wristmimic: Full-body humanoid control with wrist-guided manipulation. *ArXiv*, 2026.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. *CoRR*, abs/2507.18071, 2025.