BEYOND GLOBAL CONFLICT MINIMIZATION: EXECUTION-AWARE SKILL BLENDING FOR HUMANOID CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Composing pretrained skills offers a practical route to humanoid whole-body control, but disagreement between their action contributions is not itself a measure of execution quality. We study this distinction in a SkillBlender-based controller using mask-weighted action cancellation as a global conflict diagnostic. We introduce a jointwise execution-aware cost that couples bounded cancellation with excess joint tracking error and actuator saturation, and optimize it through PPO with a learned Lagrange multiplier. Across five paired Button training seeds, this formulation reduces mean tracking error, torque, clipping, and saturation relative to unconstrained blending, while increasing mean global conflict. Cost ablations show that minimizing cancellation alone does not reproduce these execution improvements. Re-evaluation of three Transfer training seeds at 80,000 iterations yields lower mean execution errors, but a controlled comparison of two fixed seed-33 policies over 20 matched evaluation episodes reveals substantially worse object placement and reversals in several execution metrics. These findings support separating conflict, physical execution, and task progress in the evaluation of skill composition. They do not establish universal superiority, constraint satisfaction, or real-robot safety. We provide an implementation audit and a reset-aware evaluation protocol to make these distinctions explicit.

1 INTRODUCTION

Humanoid robots must coordinate locomotion, balance, and manipulation through a shared body. Reinforcement learning has enabled increasingly capable humanoid locomotion and whole-body control in the real world (Radosavovic et al., 2023; He et al., 2024). However, training a separate controller for every combination of skills can be expensive, and a controller that solves one task need not provide reusable components for another. Skill composition addresses this problem by reusing pretrained behaviors and learning how to combine their contributions. Compositional policies and recent humanoid skill-blending systems demonstrate the promise of this approach (Peng et al., 2019; Kuang et al., 2025).

Reusing skills introduces a control problem: multiple primitives may request different motions at the same joint. A natural response is to penalize disagreement between their contributions. Yet an action-space conflict statistic is separated from physical execution by blending, action delay, clipping, position targets, feedback control, and actuator limits. Large opposing contributions may cancel to a modest command. Conversely, a single unopposed contribution can produce a difficult target or saturate an actuator. Thus, reducing global conflict does not, by definition, guarantee lower tracking error, smoother targets, or less actuator clipping. Whether it improves task performance is a further empirical question.

This distinction matters for deployment-oriented research. Training rewards summarize task incentives, while tracking errors and actuator behavior describe different aspects of what a controller actually asks the robot to do. Dynamics randomization can improve transfer from simulation (Peng et al., 2018), but it does not eliminate the need to inspect these quantities. Similarly, constrained reinforcement learning provides a way to express competing objectives (Achiam et al., 2017; Stooke

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.

et al., 2020); choosing a cost still requires care. A low learned cost should not be interpreted as a general safety certificate (Brunke et al., 2021).

We investigate these issues in a SkillBlender-based humanoid controller. Our global conflict diagnostic measures mask-weighted action cancellation. Our execution-aware cost multiplies bounded cancellation and an execution-error signal *at each joint*, before averaging over the controlled group. The execution signal combines excess target tracking error and jointwise actuator saturation. This design emphasizes situations where conflicting primitive contributions coincide with difficult physical execution. We compare it with unconstrained blending, a fixed-weight execution-gated cancellation penalty, and cancellation-only and execution-only costs.

The results support a narrower and more useful claim than universal conflict minimization. In five paired Button seeds, the proposed cost improves several execution metrics even though global cancellation increases. An archived Box cohort provides additional evidence of lower mean execution errors. However, Transfer exposes important limits. Three-seed aggregate re-evaluation suggests execution improvements, whereas a matched evaluation of one fixed policy pair shows worse task progress and worse values of several execution metrics. We report both observations rather than treating one as a substitute for the other.

Our contributions are threefold: (i) a jointwise execution-aware cost for studying conflict in skill blending; (ii) multi-task comparisons and cost ablations that separate cancellation, execution, and task progress; and (iii) an implementation and measurement audit, including a matched Transfer case study, that identifies the scope and limitations of the empirical conclusions. The optimizer is an application of PPO-Lagrangian, not a new constrained optimization algorithm. All results are from simulation.

2 Related work

Skill composition and humanoid control. Multiplicative compositional policies combine reusable primitives to support hierarchical control (Peng et al., 2019). SkillBlender applies skill blending to humanoid whole-body loco-manipulation (Kuang et al., 2025). Our study uses this latter setting and focuses on how blending should be evaluated and regularized when primitive contributions interact. It complements advances in humanoid locomotion and teleoperation (Radosavovic et al., 2023; He et al., 2024); it does not introduce a new primitive library or demonstrate a new hardware capability.

Constrained policy learning. PPO offers a practical clipped-surrogate policy update (Schulman et al., 2017). Constrained Policy Optimization develops a trust-region approach to constrained reinforcement learning (Achiam et al., 2017), while Lagrangian methods adapt the weight of a cost during training. The dynamics of this adaptation can themselves affect learning (Stooke et al., 2020). We use a PPO reward objective, a cost critic, and projected dual ascent. Our contribution concerns the cost signal and its empirical behavior; theoretical guarantees for other constrained algorithms do not transfer to this implementation.

Execution quality and evaluation. Safe learning in robotics spans constraint design, uncertainty, and control-system considerations (Brunke et al., 2021). Simulation-to-real transfer additionally depends on modeling and robustness (Peng et al., 2018). We therefore treat tracking, clipping, and saturation as execution diagnostics, not as complete definitions of safety. Evaluation in reinforcement learning is sensitive to seeds and aggregation (Henderson et al., 2018; Agarwal et al., 2021). We distinguish training-seed variation from repeated evaluation of fixed policies, disclose legacy metric limitations, and avoid interpreting thousands of correlated timesteps as independent replications.

3 Skill blending and physical execution

Let $a_{t,k,j}$ be primitive k's proposed action for joint j at control step t, and $m_{t,k,j}$ its blend mask. The primitive networks are frozen while the blending policy learns their masks. Define the contribution

$z_{t,k,j} = m_{t,k,j} a_{t,k,j}$ and the blended policy action

$$a^{\mathrm{pol}}_{t,j} = \sum_{k=1}^{K} z_{t,k,j}. \tag{1}$$

The audited controller has 19 action dimensions and uses pretrained walking and reaching skills. We define raw joint cancellation and its global diagnostic as

$$c^{\mathrm{raw}}_{t,j} = \sum_{k} |z_{t,k,j}| - \left| \sum_{k} z_{t,k,j} \right|, \qquad G_t = \frac{1}{19} \sum_{j=1}^{19} c^{\mathrm{raw}}_{t,j}. \tag{2}$$

The triangle inequality ensures nonnegativity. This definition measures cancellation of signed contributions; it is *not* a pairwise distance between primitive actions. Unless otherwise specified, "global conflict" in this paper means G_t.

A simple example illustrates why cancellation is not an execution metric. Contributions $z_1 = x$ and $z_2 = -x$ yield cancellation $2|x|$ but a zero blended action. A single nonzero contribution yields zero cancellation, regardless of its magnitude. Neither example determines the robot's tracking response without the controller state and dynamics. Consequently, Eq. 2 alone cannot order policies by physical execution quality.

From blended actions to torques. The audited Transfer pipeline applies randomized action delay, action clipping, target construction, and PD control:

$$a^{\mathrm{delay}}_t = (1 - d_t) a^{\mathrm{pol}}_t + d_t a^{\mathrm{exec}}_{t-1}, \qquad a^{\mathrm{exec}}_t = \mathrm{clip}(a^{\mathrm{delay}}_t), \tag{3}$$

$$q^{\mathrm{des}}_t = q^{\mathrm{default}} + s_a a^{\mathrm{exec}}_t, \qquad \tau^{\mathrm{raw}}_t = K_p(q^{\mathrm{des}}_t - q_t) - K_d \dot{q}_t, \tag{4}$$

$$\tau^{\mathrm{app}}_t = \mathrm{clip}(\tau^{\mathrm{raw}}_t, -\tau^{\max}, \tau^{\max}). \tag{5}$$

Here d_t is sampled uniformly in $[0, 1]$ for each environment and broadcast across joints, and $s_a = 0.25$ in the audited Transfer configuration. A discrepancy between policy and executed actions is therefore expected and is not, by itself, a logging error.

Execution diagnostics. For a joint set $\mathcal{D}$, tracking is the mean of $|q^{\mathrm{des}} - q|$; torque is mean absolute applied torque; clipping is mean $|\tau^{\mathrm{raw}} - \tau^{\mathrm{app}}|$; and saturation is the fraction of joint samples with clipping greater than 10^{-6}. Target jump is the mean absolute change in q^{des} between consecutive control steps *within the same episode*. First steps following a reset are excluded. Tracking and jump are in radians; torque and clipping are in N m. Audited Transfer diagnostics use all 19 joints, whereas the learning cost acts on the right arm. The two scopes must not be conflated.

4 Execution-aware constrained learning

We first bound cancellation separately at each joint:

$$C_{t,j} = \frac{c^{\mathrm{raw}}_{t,j}}{1 + c^{\mathrm{raw}}_{t,j}}. \tag{6}$$

Let $e_{t,j} = |q^{\mathrm{des}}_{t,j} - q_{t,j}|$ and $b_{t,j} = |\tau^{\mathrm{raw}}_{t,j} - \tau^{\mathrm{app}}_{t,j}|$. The execution signal is

$$E^{\mathrm{track}}_{t,j} = \mathrm{clip}\left(\frac{\max(e_{t,j} - \tau_q, 0)}{\tau_q + 10^{-8}}, 0, 1 \right), \tag{7}$$

$$E^{\mathrm{sat}}_{t,j} = \mathbf{1}[b_{t,j} > 10^{-6}], \qquad E_{t,j} = \tfrac{1}{2}(E^{\mathrm{track}}_{t,j} + E^{\mathrm{sat}}_{t,j}). \tag{8}$$

The frozen configuration uses $\tau_q = 0.10940390080213547$. Saturation is jointwise, rather than a single environment-level event broadcast to every joint. For the right-arm group $\mathcal{J}$ (zero-based indices 15–18), we compare

$$\ell^C_t = \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} C_{t,j}, \quad \ell^E_t = \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} E_{t,j}, \quad \ell^{C \times E}_t = \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} C_{t,j} E_{t,j}. \tag{9}$$

The order of operations matters. Bounding after averaging differs from averaging bounded values, and the mean of jointwise products differs from the product of group means. The proposed cost specifically measures jointwise co-occurrence.

Policy and dual updates. The nominal motivation is to maximize task return subject to an expected cost budget d. In implementation, we minimize the PPO reward loss plus a multiplier-weighted clipped cost surrogate and value losses. With probability ratio $r_t(\theta)$ and cost advantage $\hat{A}_t^C$, the cost surrogate is

$$L_C(\theta) = \mathbb{E}_t\Big[\max\Big(r_t(\theta)\hat{A}_t^C, \mathrm{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t^C\Big)\Big]. \tag{10}$$

A separate cost critic supplies discounted cost estimates. The dual update is

$$\lambda \leftarrow \mathrm{clip}\left(\lambda + \eta_\lambda(\widehat{J}_C^{\mathrm{store}} - d), 0, \lambda_{\max}\right). \tag{11}$$

Importantly, $\widehat{J}_C^{\mathrm{store}}$ is the mean of *stored* rollout costs. The audited code adds a discounted cost-value bootstrap at timeouts before storing those costs. It is therefore not necessarily the mean of the immediate physical cost used in calibration. We report this implementation faithfully and do not claim that Eq. 11 enforces an immediate-cost constraint. Default dual settings in the frozen implementation are initialization zero, learning rate 0.01, and cap 10; run-specific overrides are not established by the default values alone.

The recorded $C \times E$ budget is $d_{C\times E} = 0.0105962552$, calibrated from the reference mean of five N1 rollouts. The recorded E-only budget is $d_E = 0.0673833951$. These different costs have different scales; their numerical budgets are not interchangeable. Our experiments compare configured methods, rather than a complete matched-performance budget frontier.

Baselines and scope. H0 is unconstrained skill blending. N1 is a fixed-weight, execution-gated cancellation auxiliary loss, with coefficient 0.001:

$$L_{\mathrm{N1}} = \frac{\sum_{t,j} g_{t,j} c_{t,j}^{\mathrm{raw}}}{\sum_{t,j} g_{t,j} + 10^{-8}}, \qquad g_{t,j} = \mathbf{1}[e_{t,j} > \tau_q \text{ or } b_{t,j} > 10^{-6}]. \tag{12}$$

The gate is detached. Although N1's configured group is both arms, the frozen environment masks its execution gate to the right arm; the effective nonzero terms are consequently right-arm terms. N1 must not be relabeled as a C-only cost. We use "PPO-Lagrangian" or "$C\times E$" for the same proposed variant in the corresponding experiments.

The product cost has an explicit blind spot: if $C_{t,j} = 0$, it assigns zero cost even when $E_{t,j}$ is high. It cannot bound unconditional tracking error or saturation. E-only is therefore a substantive comparator, and a benefit over cancellation-only is insufficient to establish superiority over direct execution regularization.

5 Experimental setup

We distinguish three sources of evidence. First, archived Button results include five paired training seeds (26–30) for H0, N1, and $C \times E$, together with C-only and E-only ablations. The $C \times E$ ablation row reuses the same five policies; it is not an independent replication. Archived Box results compare three seeds (31–33). Second, an audited Transfer re-evaluation compares three training seeds (32–34), using checkpoints at 80,000 iterations and 2,500 logged transitions per policy. These checkpoints are not asserted to be converged or equivalent to the configured 100,001-iteration endpoint. Third, a matched Transfer case study evaluates only the two seed-33 policies over evaluation seeds 1001–1020, for 250 steps each.

Measurement and aggregation. The unified audit records policy contributions, executed actions, targets, positions, and torques under one transition index, before resets. Environment, episode, and timestep identifiers support reset-aware comparisons. In a 300-step reconstruction check, the masked primitive sum exactly reproduces the logged policy action. Across Transfer traces, episode IDs span 10–11 episodes per policy; valid within-episode target differences number 2,489–2,490 rather than 2,500. We calculate each policy's time average and report the mean and sample standard deviation across training seeds. For the matched case study, standard deviations describe evaluation episodes of *fixed policies*, not uncertainty across training runs.

Archived Button/Box summaries are retained under their original evaluation protocol and are not pooled with the corrected all-joint Transfer data. In particular, their target-jump summaries include

Table 1: Button, five paired training seeds (26–30), archived protocol. Mean ± reported standard deviation across seeds. Lower execution metrics are desirable; cancellation has no presumed monotonic relation to task quality.

Metric	H0	N1	$C \times E$
Global cancellation	$.2942 \pm .1266$	$.2882 \pm .1432$	$.3721 \pm .1479$
Tracking (rad)	$.0818 \pm .0233$	$.0829 \pm .0124$	$.0513 \pm .0061$
Torque (N m)	5.954 ± 1.642	$5.586 \pm .901$	$3.809 \pm .560$
Clipping (N m)	$.4138 \pm .5053$	$.5781 \pm .5045$	$.0198 \pm .0180$
Saturation fraction	$.0231 \pm .0183$	$.0280 \pm .0186$	$.0023 \pm .0022$

reset boundaries. We omit these legacy jumps from the main execution tables and document them in the appendix. This prevents a change in logging semantics from appearing to be an algorithmic improvement.

Task progress. For Transfer, object placement error is $e_t^{\text{box}} = \frac{1}{3}\sum_{u=1}^{3} |p_{t,u}^{\text{box}} - p_{t,u}^{\text{goal}}|$, in meters. The corresponding reward component is $\exp(-4e_t^{\text{box}})$. We report final and minimum error rather than inventing a success threshold. The audited termination code distinguishes abnormal-contact termination and timeout, not a task-success flag. In matched evaluation, initial box positions, goals, and action-delay sequences are controlled across the policy pair. Other unrecorded sources of variation are not assumed to be matched.

Inference. Our main tables are descriptive comparisons. Training-seed counts are small. For example, five same-direction nonzero paired differences have a two-sided exact Wilcoxon signed-rank p-value of 0.0625, so consistency across five seeds does not establish significance at 0.05. We do not present formal Pearson/Spearman correlations from the corrected protocol: the planned seed-, episode-, and timestep-level analysis, with dependence-aware uncertainty, is not complete. Method-level mean changes are not a substitute for those correlations.

6 Results

6.1 Button: better execution can coexist with higher cancellation

Table 1 compares the paired Button cohort. Relative to H0, $C \times E$ decreases mean tracking error by 37.3%, mean applied torque by 36.0%, clipping by 95.2%, and saturation by 90.0%, while mean global cancellation increases by 26.5%. Tracking, torque, clipping, and saturation are lower in all five paired seeds; cancellation is higher in four. These results contradict the use of lower global cancellation as a universal ordering of controllers by execution quality. They do not establish that increasing cancellation causes the improvements.

N1 has slightly lower mean cancellation than H0, but higher tracking error, clipping, and saturation. Its execution gate and fixed auxiliary weight do not reproduce the $C \times E$ results in this cohort. Task evidence is less decisive: the archived Button task metric changes from 0.0745 to 0.0664, but improves in only two of five paired seeds. We do not infer consistent task improvement from that group mean or equate it with a success rate.

6.2 Cost ablations: cancellation alone is insufficient

Table 2 compares the three constrained cost definitions. C-only achieves the lowest cancellation but has the highest tracking error, torque, clipping, and saturation among these variants. E-only and $C \times E$ both substantially improve mean execution metrics relative to C-only. The $C \times E$ variant has lower mean tracking, clipping, and saturation than E-only, while E-only has slightly lower torque. In particular, the mean clipping reduction relative to E-only is 49.3%, and the saturation reduction is 42.8%. With five seeds and no completed matched-task budget sweep, these descriptive improvements do not establish a general advantage of multiplicative coupling. The evidence supports evaluating execution explicitly, while leaving the best coupling and tradeoff unresolved.

Table 2: Button cost ablations, seeds 26–30. Mean ± reported standard deviation. The $C{\times}E$ policies are the same as in Table 1.

Metric	C-only	E-only	$C \times E$
Global cancellation	$.1435 \pm .0525$	$.2624 \pm .1433$	$.3721 \pm .1479$
Tracking (rad)	$.0837 \pm .0163$	$.0546 \pm .0261$	$.0513 \pm .0061$
Torque (N m)	$6.016 \pm .798$	3.787 ± 1.458	$3.809 \pm .560$
Clipping (N m)	$.3673 \pm .3309$	$.0391 \pm .0612$	$.0198 \pm .0180$
Saturation fraction	$.0235 \pm .0163$	$.0040 \pm .0065$	$.0023 \pm .0022$

Table 3: Audited Transfer re-evaluation at 80,000 iterations. Mean ± sample standard deviation over three training seeds (32–34); 2,500 transitions per policy. Execution metrics cover all 19 joints, while the learning cost uses the right arm.

Metric	H0	$C \times E$
Global cancellation	$.1384 \pm .0570$	$.1566 \pm .0260$
Tracking (rad)	$.1601 \pm .0492$	$.1158 \pm .0015$
Within-episode jump (rad)	$.0255 \pm .0028$	$.0238 \pm .0058$
Torque (N m)	18.138 ± 3.319	$14.908 \pm .453$
Clipping (N m)	1.3366 ± 1.0482	$.3738 \pm .1836$
Saturation fraction	$.0689 \pm .0609$	$.0270 \pm .0046$

6.3 Box: additional aggregate evidence

In the archived three-seed Box comparison, H0 versus $C \times E$ mean tracking error is 0.0608 versus 0.0415 rad, mean torque is 5.106 versus 3.646 N m, clipping is 0.6154 versus 0.2503 N m, and saturation is 0.0323 versus 0.0094. Here cancellation also decreases, from 0.1492 to 0.0221. Thus, increased cancellation is not a necessary condition for better aggregate execution. This cohort provides a second task context, but does not establish task-success improvements or uniform seedwise dominance. Full reported dispersion appears in Appendix B.

6.4 Transfer: aggregate execution and a matched failure case

The corrected three-seed Transfer evaluation in Table 3 uses the same all-19-joint metric definitions for both variants and excludes resets from target jumps. At 80,000 iterations, $C{\times}E$ has 27.7% lower mean tracking error, 17.8% lower torque, 72.0% lower clipping, and 60.8% lower saturation, with 13.2% higher cancellation. Mean target jump decreases by 6.6%. These are checkpoint-specific averages with substantial seed variation, particularly for H0 clipping and saturation. They cannot establish the result of longer training.

The matched seed-33 case study answers a different question: how do two particular learned policies behave under controlled evaluation conditions? Over 20 matched episodes, H0 achieves lower final and minimum object error in every pair. Mean final error is 0.0100 m for H0 and 0.1909 m for $C \times E$ (Table 4). All 40 rollouts reach the 250-step evaluation timeout without abnormal-contact termination. A timeout is not evidence of either task success or failure; the continuous object error supplies the task comparison.

This fixed pair also reverses several aggregate execution comparisons: $C \times E$ has higher tracking error, clipping, and saturation, despite slightly lower torque and target jump. It would therefore be inaccurate to describe the matched experiment as improved execution at the expense of task progress. It is a failure case for both task error and several execution diagnostics. The three-seed table and the fixed-pair table estimate different quantities and should not be combined into a single sample count.

Figure 1 shows the temporal structure. Both policies start at the same mean object error. H0 begins transporting the object and reaches low error, while $C \times E$ retains a large mean object error for much of the rollout. Wrist proximity alone does not explain successful transport, and the clipping trajectory reveals an additional execution difference during interaction. This pattern is consistent with stalled or ineffective object interaction. The controlled initial box state, goal, and delay se-

Table 4: Matched Transfer case study: two fixed seed-33 policies, 20 paired evaluation seeds (1001–1020), 250 steps per rollout. Standard deviations are across evaluation episodes, not training seeds.

Metric	H0	$C \times E$
Final object error (m)	$.0100 \pm .0056$	$.1909 \pm .1438$
Minimum object error (m)	$.0075 \pm .0044$	$.1896 \pm .1446$
Tracking (rad)	$.1127 \pm .0042$	$.1206 \pm .0071$
Within-episode jump (rad)	$.0272 \pm .0028$	$.0262 \pm .0033$
Torque (N m)	$15.194 \pm .604$	$14.911 \pm .891$
Clipping (N m)	$.2433 \pm .0551$	$.5686 \pm .3915$
Saturation fraction	$.0176 \pm .0052$	$.0299 \pm .0082$

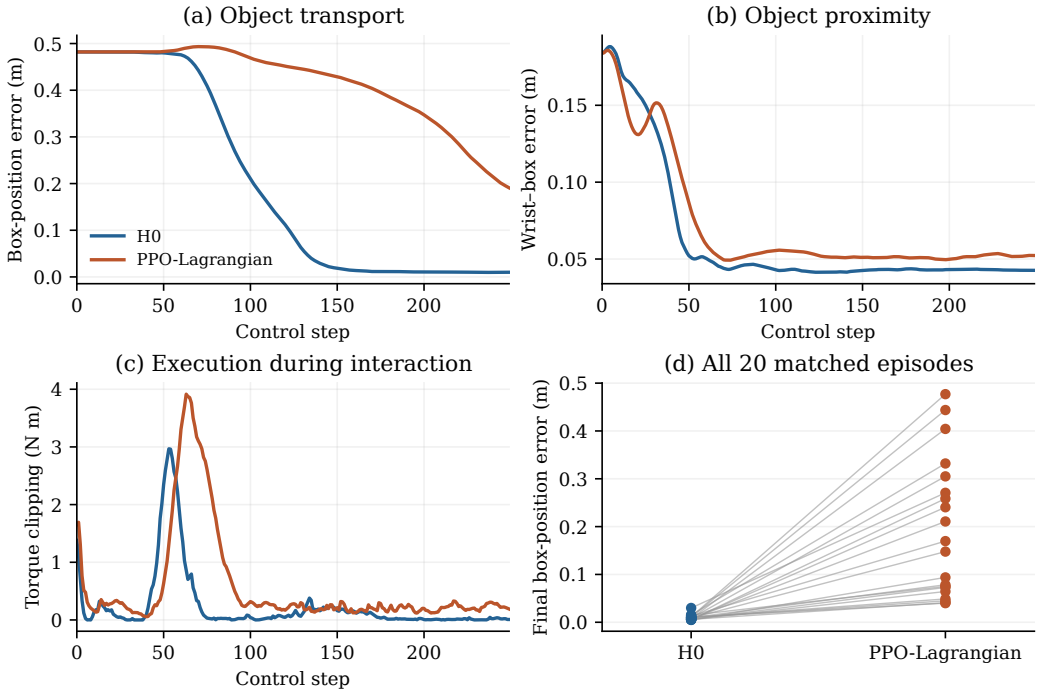

Figure 1: Matched Transfer evaluation of the fixed seed-33 policy pair. The first three panels show means over 20 matched episodes, without confidence bands; the final panel displays all paired final object errors. The task-progress gap is substantial even though the policies share the controlled initial box positions, goals, and action-delay sequences. Curves do not represent variability across trained policies.

quences reduce specific evaluation confounds, but do not identify the training cost as the sole cause. Establishing that causal mechanism requires further controlled interventions and additional training seeds.

7 Discussion and Limitations

Our evidence supports separating three questions: do primitive contributions cancel, can the controller execute the resulting targets, and does the robot make task progress? Button demonstrates that better execution metrics can coexist with higher global cancellation. Box demonstrates that both can decrease. Transfer demonstrates that favorable averages across a small training cohort neither guarantee a particular policy's execution quality nor establish good object transport. These observations motivate evaluating the quantities separately rather than combining them into an unvalidated deployment score.

Several limitations determine the scope of the paper. First, the product cost ignores execution errors when cancellation is zero and applies only to the right arm, while whole-body performance depends on additional joints and contacts. Second, the dual statistic includes timeout bootstrapping and is not identical to the immediate-cost quantity used for calibration. Constraint feasibility is not established. Third, legacy Button/Box summaries have not all been reprocessed under the corrected reset-aware protocol. Fourth, training-seed counts are small and the matched Transfer study concerns one selected pair, motivated by an observed discrepancy rather than a representative random sample of policy pairs.

Fifth, a completed fixed-penalty versus budget tradeoff curve at matched task performance is unavailable. Recent Button sweep outputs lack required task-position fields and contain multiple episodes despite the intended single-episode protocol; they are excluded. Sixth, formal conflict–execution correlations with dependence-aware confidence intervals remain uncomputed under the unified protocol. Finally, there are no G1 hardware results, no demonstrated robustness to deployment perturbations, and no safety guarantee. These omissions limit deployment claims, but do not erase the documented distinction between action cancellation and physical execution.

8 Conclusion

We studied execution-aware skill blending through jointwise cancellation and execution signals. The proposed $C \times E$ cost improves several mean execution metrics in Button and in aggregate Box and Transfer comparisons, while cancellation need not decrease. A matched Transfer case study exposes substantial task degradation and reversals in several execution comparisons. The central conclusion is therefore an evaluation principle supported by an empirical method study: global cancellation, physical execution, and task progress must be measured separately. The current evidence motivates further task-aware cost design and controlled tradeoff studies, rather than a claim that one conflict metric or constraint is sufficient for deployable humanoid control.

Reproducibility statement

The paper specifies the audited cancellation and execution formulas, the joint groups, the action pipeline, the available budgets, and the distinction between training seeds and evaluation episodes. Appendix A records the frozen Transfer configuration and audit procedure. Tables distinguish archived summaries from recomputed traces. The source package includes the matched-evaluation data and figure script. Complete retraining additionally requires the underlying simulator, primitive checkpoints, and run-specific configurations; these are not all contained in the manuscript source package.

Ethics statement

The reported experiments use simulated humanoid control and contain no human-subject study. Lower tracking error or actuator saturation should not be taken as evidence that a policy is safe around people. Hardware validation and a broader assessment of contact behavior remain necessary before deployment.

AI use statement

We used ChatGPT to assist in assessing the feasibility of research directions, refining experimental designs, identifying relevant literature, reviewing code, summarizing experimental results, preparing figures, and polishing the manuscript. The authors take full responsibility for the accuracy and integrity of the work.

References

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceed-*

ings of Machine Learning Research, pp. 22–31, 2017. URL https://proceedings.mlr.press/v70/achiam17a.html.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. *Advances in Neural Information Processing Systems*, 34, 2021. URL https://arxiv.org/abs/2108.13264.

Lukas Brunke, Melissa Greeff, Adam W. Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P. Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *arXiv preprint arXiv:2108.06266*, 2021. URL https://arxiv.org/abs/2108.06266.

Tairan He, Zhengyi Luo, Wenli Xiao, Chong Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Learning human-to-humanoid real-time whole-body teleoperation. *arXiv preprint arXiv:2403.04436*, 2024. URL https://arxiv.org/abs/2403.04436.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. URL https://arxiv.org/abs/1709.06560.

Yuxuan Kuang, Haoran Geng, Amine Elhafsi, Tan-Dzung Do, Pieter Abbeel, Jitendra Malik, Marco Pavone, and Yue Wang. Skillblender: Towards versatile humanoid whole-body loco-manipulation via skill blending. *arXiv preprint arXiv:2506.09366*, 2025. URL https://arxiv.org/abs/2506.09366.

Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. *arXiv preprint arXiv:1710.06537*, 2018. URL https://arxiv.org/abs/1710.06537.

Xue Bin Peng, Michael Chang, Grace Zhang, Pieter Abbeel, and Sergey Levine. MCP: Learning composable hierarchical control with multiplicative compositional policies. *arXiv preprint arXiv:1905.09808*, 2019. URL https://arxiv.org/abs/1905.09808.

Ilija Radosavovic, Tete Xiao, Bike Zhang, Trevor Darrell, Jitendra Malik, and Koushil Sreenath. Real-world humanoid locomotion with reinforcement learning. *arXiv preprint arXiv:2303.03381*, 2023. URL https://arxiv.org/abs/2303.03381.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by PID lagrangian methods. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9133–9143, 2020. URL https://proceedings.mlr.press/v119/stooke20a.html.

A AUDITED CONFIGURATION AND MEASUREMENT PROTOCOL

Table 5 records values from the frozen Transfer configuration. They describe the inspected configuration rather than proving that every archived run used identical overrides. In particular, the 250-step matched evaluation horizon differs from the configured eight-second training episode. The evaluated checkpoint is iteration 80,000, not the configured training endpoint.

Table 5: Frozen Transfer configuration.

Setting	Value
Physics step / control decimation	0.001 s / 10
Control frequency	100 Hz
Parallel training environments	4,096
Configured episode duration	8 s
Action dimensions / action scale	19 / 0.25
Actor hidden widths	512, 256, 128
Critic hidden widths	768, 256, 128
Initial policy learning rate	10^{-5} (adaptive KL schedule)
Learning epochs / minibatches	2 / 4
Discount / GAE parameter	0.994 / 0.9
Entropy coefficient	0.001
Rollout steps per environment	60
Configured maximum iterations	100,001
Evaluated checkpoint	80,000
Matched evaluation horizon	250 control steps

The configuration assigns PD stiffness 200 to hip yaw/roll/pitch, 300 to knees and torso, 40 to ankles, and 100 to shoulders and elbows. Corresponding damping is 5 for hips, 6 for knees and torso, and 2 for ankles, shoulders, and elbows. These parameters explain why identical action-space cancellation need not translate into identical torque behavior across joints.

Audit procedure. The frozen files cover the hierarchical actor, PPO update, robot controller, evaluator, analysis script, and Transfer task configuration. We distinguish primitive actions, masked contributions, blended policy actions, delayed/clipped executed actions, position targets, and raw/applied torque. Reconstructing the blend from the logged contributions checks one identity, not the entire physical simulation. The logger uses a common transition index and records observations before reset replacement. Valid target-jump samples require equal consecutive episode IDs within the same environment. This prevents reset changes from being interpreted as control discontinuities.

Budget interpretation. The calibrated $C \times E$ budget is a reference mean from five N1 rollouts. It is a design choice rather than a measured physical safety threshold. The stored training cost can receive timeout bootstrapping, whereas an evaluation trace reports immediate physical diagnostics. A future implementation intended to enforce an immediate-cost budget should explicitly align these statistics and report cost satisfaction separately. No such alignment or feasibility result is claimed here. The precise C-only budget and all run overrides are not recoverable from the verified summary used in this manuscript; this limits exact retraining from the paper alone.

Aggregation and dependence. For each re-evaluated Transfer policy, we average immediate diagnostics over its logged transitions and all joints. Seed-level averages supply Table 3; a long trajectory is not a large independent sample of trained policies. Matched evaluation uses identical recorded evaluation conditions for two fixed policies. Its sample dispersion describes sensitivity to those conditions only. A future formal correlation analysis should report within-policy episode/timestep associations separately from between-policy or between-seed associations, exclude reset transitions where appropriate, and use episode- or seed-level resampling rather than treating adjacent timesteps as independent. These are analysis requirements, not completed results.

B ADDITIONAL ARCHIVED RESULTS

The archived summaries retain useful comparisons, but their metric provenance is weaker than that of the re-evaluated Transfer traces. We therefore do not pool them with the all-joint audited metrics. Table 6 reports the Box cohort with its original dispersion. Table 7 documents reset-inclusive jumps for transparency; these values are not equivalent to the reset-aware jumps in the main Transfer tables.

Table 6: Archived Box comparison, three training seeds (31–33). Mean ± reported standard deviation.

Metric	H0	$C \times E$
Global cancellation	$.1492 \pm .0125$	$.0221 \pm .0157$
Tracking (rad)	$.0608 \pm .0254$	$.0415 \pm .0153$
Torque (N m)	5.106 ± 1.754	3.646 ± 1.208
Clipping (N m)	$.6154 \pm .8108$	$.2503 \pm .4093$
Saturation fraction	$.0323 \pm .0367$	$.0094 \pm .0114$

Table 7: Legacy reset-inclusive target jump (rad). These summaries are retained for provenance and are not used to compare against corrected Transfer jump values.

Task	Variant	Mean ± reported standard deviation
Button	H0	$.04231 \pm .00841$
Button	N1	$.04322 \pm .00562$
Button	C-only	$.04721 \pm .00934$
Button	E-only	$.03131 \pm .01191$
Button	$C \times E$	$.02880 \pm .00577$
Box	H0	$.01709 \pm .00116$
Box	$C \times E$	$.01439 \pm .00154$

The archived Button task metric is 0.074467 ± 0.026858 for H0, 0.066745 ± 0.023535 for N1, and 0.066446 ± 0.027134 for $C \times E$. The available scalar summaries do not justify a new success definition. In the cost ablation, the same metric is 0.062093 ± 0.012397 for C-only and 0.085308 ± 0.036674 for E-only. These task summaries reinforce the need for matched-task comparisons rather than ranking methods solely by execution diagnostics.

C EXCLUDED EXPERIMENTS AND LIMITS OF INTERPRETATION

A recent Button sweep includes fixed execution penalties with coefficients 0.0003, 0.001, and 0.003, and E-only and $C \times E$ budgets at 0.75, 1, and 1.25 times their reference values. The available evaluation outputs are not a valid task-performance frontier: required wrist and button-goal fields are absent, and the intended stop-after-done behavior was not implemented, producing multiple episodes within capped rollouts. We exclude these outputs from quantitative claims. Completion of training alone is insufficient to validate the evaluation.

Neither the archived summaries nor the corrected Transfer comparisons provide the requested full Pearson/Spearman analysis between global conflict and tracking, jump, torque, clipping, and saturation at seed, episode, and timestep levels. In particular, a positive difference in mean cancellation coupled with a negative difference in mean clipping is not a correlation coefficient. Previously row-aligned quantities with uncertain transition correspondence are not used. The present paper is an empirical comparison of methods and an audit of measurements, not a completed statistical validation of global conflict as a predictive signal.

No G1 experiment is reported. The matched Transfer case study does not establish a general causal law. Task-progress-aware constraints and replication across trained policy pairs remain future work.