Extended Abstract

Motivation. This project is motivated by pose-based volleyball coaching, where video-derived keypoints could evaluate approach timing, jump mechanics, torso position, and arm-swing form. A natural RL formulation is to treat expert movement as a demonstration and convert pose similarity into a dense reward. However, the same pose signal that makes this feedback interpretable is also error-prone. Fast motion can cause jitter, occlusion can remove keypoints, and camera viewpoint can introduce systematic bias. If corrupted keypoints define the reward optimized by PPO, the agent may optimize a signal that no longer matches the intended movement objective. This motivates the central question: which pose-estimation errors are most harmful to reward alignment, and when do alignment failures lead to downstream policy errors?

Method. I study this question in MuJoCo by isolating corruption in the reward channel. The policy observes the clean simulator state, while only the reward keypoints are corrupted. I collect expert reference trajectories, extract root-relative simulator body keypoints, and define a dense imitation reward from policy-expert keypoint distance. PPO policies are trained under structured corruptions matching common pose-estimation errors, including Gaussian jitter, zero-imputed dropout, previous-frame hold dropout, masked dropout, and systematic bias. Policies are evaluated under clean keypoints using pose return, environment return, episode length, and final keypoint MSE.

Implementation. The main controlled experiment uses Walker2d-v5 because its clean pose reward is learnable while still sensitive to reward corruption. I extend the pipeline to Hopper-v5 as a second morphology and to Humanoid-v5 as a full-body stress test with torso, pelvis, legs, feet, and upper-arm keypoints. The Humanoid extension adds arm-specific bias to connect more directly to full-body sports feedback, where upper-limb tracking errors matter for technique evaluation. For each environment, I collect demonstrations from trained default-reward PPO experts and train pose-reward PPO agents under clean and corrupted rewards.

Results. The main finding is that missing-keypoint handling changes the reward ordering optimized by PPO, not just the reward scale. In a dropout-rate sweep across the three environments, zero-imputation rapidly collapses Spearman alignment between the clean oracle reward and corrupted training reward, while masked scoring and hold-last imputation preserve higher alignment. In Walker2d, this alignment failure corresponds to worse downstream behavior. Zero-imputed dropout produces higher clean-evaluation keypoint error, while masked scoring improves clean-evaluation return and reduces keypoint MSE relative to zero-imputation. Cross-environment results qualify this pattern. In Hopper, several corrupted rewards still reach stable 1000-step behavior despite low alignment, suggesting that simpler morphologies can tolerate reward distortion. In Humanoid, high-alignment rewards still produce short-horizon policies, showing that full-body pose-reward imitation can remain difficult even when the reward is well aligned.

Discussion. These results show that reward-task alignment is useful but incomplete as a diagnostic for pose-derived reward reliability. Low alignment reveals when corrupted keypoints make the reward rank behavior differently from the clean imitation objective, and this predicts downstream failure in Walker2d. However, Hopper and Humanoid show that alignment alone does not determine final PPO performance. Morphology, optimization difficulty, reference complexity, and reward scale also affect whether corruption harms learning. The main practical implication is that missing keypoints should not be replaced with arbitrary coordinates such as zero. Masked or confidence-weighted scoring is more robust because it avoids penalizing the agent for pose-estimator artifacts.

Conclusion. This project studies pose-derived reward corruption in demonstration-guided PPO, motivated by pose-based coaching feedback. The core contribution is an analysis of a reward-design failure mode that arises when dense imitation rewards are computed from unreliable keypoints. Across three MuJoCo morphologies, zero-imputed dropout consistently damages reward alignment, while masked scoring better preserves the intended reward structure. Robust pose-based reinforcement learning therefore requires reliable reward construction and sufficient optimization machinery for the underlying control problem.

Pose Under Pressure: Robustness of Pose-Derived Dense Rewards in Demonstration-Guided Reinforcement Learning

Joseph Dehoney
Stanford University
jdehoney@stanford.edu

Abstract

Pose-derived dense rewards are a natural way to train reinforcement learning agents from demonstrations, but they depend on the reliability of the keypoints used to compute the reward. This project is motivated by pose-based volleyball coaching, where keypoints could evaluate approach timing, jump mechanics, torso position, and arm-swing form. I study this problem in MuJoCo by isolating corruption in the reward channel. The policy observes the clean simulator state, while only the keypoints used in the scalar imitation reward are corrupted. Across Walker2d-v5, Hopper-v5, and Humanoid-v5, I evaluate Gaussian jitter, zero-imputed dropout, previous-frame hold dropout, masked dropout, and systematic keypoint bias. The results show that missing-keypoint handling is critical. Zero-imputed dropout consistently damages reward alignment across morphologies, while masked and hold-last scoring preserve more reward structure. In Walker2d, low reward alignment corresponds to worse downstream PPO behavior and masked scoring improves clean-evaluation performance, although the mitigation result is noisy across seeds. In Hopper, low alignment does not necessarily cause failure, and in Humanoid, high alignment does not guarantee successful imitation. These results show that reward alignment is a useful diagnostic for corrupted pose rewards, but final policy quality also depends on morphology and optimization difficulty.

1 Introduction

Pose-based coaching systems aim to evaluate movement quality from video. In volleyball, video-derived keypoints could score approach timing, jump mechanics, torso position, and arm-swing form when evaluating an outside hitting attack. A natural way to connect this setting to reinforcement learning is to treat expert movement as a demonstration and convert pose similarity into a dense reward. Instead of manually designing a task reward, the agent can be rewarded for matching the keypoint trajectory of an expert.

This formulation is appealing because keypoints provide interpretable feedback, but the same pose signal is also error-prone. Fast motion can cause frame-to-frame jitter, self-occlusion can produce missing detections, and camera viewpoint can introduce systematic keypoint bias. If these corrupted keypoints are used to compute the reward optimized by PPO, the agent may optimize a signal that no longer matches the intended movement objective. The central question is therefore which pose-estimation errors most damage reward alignment and when those alignment failures lead to downstream policy errors.

This project studies that question in Walker2d-v5, Hopper-v5, and Humanoid-v5. These environments are not volleyball simulators, but they provide controlled articulated-body testbeds for isolating the reward-design question. Walker2d is the main controlled setting because its pose reward is learnable

while still sensitive to corruption. Hopper tests whether the same reward distortions matter in a simpler morphology, and Humanoid tests whether the conclusions extend to a harder full-body setting with upper-limb keypoints.

The main experimental control is that policy observations remain clean, while only the keypoints used to compute the pose reward are corrupted. This separates reward corruption from observation noise and focuses the experiments on how keypoint errors change the scalar reward optimized by PPO.

The project makes three contributions. First, it implements a pose-derived reward pipeline using expert reference trajectories and root-relative simulator body keypoints. Second, it evaluates structured reward corruptions that correspond to common pose-estimation errors. Third, it analyzes reward-task alignment using Spearman correlation between the clean oracle reward and corrupted training reward.

The central result is that missing-keypoint handling matters more than dropout rate alone. Replacing missing keypoints with zeros rapidly destroys reward alignment, while masked scoring and hold-last imputation preserve more reward structure. In Walker2d, this misalignment corresponds to worse downstream behavior, while Hopper and Humanoid show that alignment is diagnostic but not sufficient. Robust pose-based reinforcement learning therefore requires both reliable reward construction and sufficient optimization machinery.

2 Related Work

Pose-based imitation rewards. Pose-based imitation learning methods use expert motion references to train agents to reproduce demonstrated behavior. DeepMimic defines dense imitation objectives using pose, velocity, end-effector, and center-of-mass terms, allowing simulated characters to imitate reference motions (Peng et al., 2018). Adversarial Motion Priors replace hand-designed tracking objectives with learned style rewards from motion data, enabling physically simulated characters to produce natural behaviors while optimizing task objectives (Peng et al., 2021). These methods show that motion-derived rewards can support complex control, but they usually assume that the pose features used for reward computation are reliable. This project focuses on what happens when that assumption fails.

Pose estimation and missing keypoints. Pose-estimation systems often output noisy, missing, or low-confidence keypoints. OpenPose is a widely used real-time multi-person pose-estimation system that detects body, foot, hand, and face keypoints (Cao et al., 2019). More recent skeleton-processing work such as DISK directly studies missing skeleton data and uses learned imputation to recover missing keypoints from temporal and cross-keypoint structure (Rose et al., 2026). These issues are especially relevant in sports video because fast limb motion, self-occlusion, clothing, and camera angle can all affect keypoint quality. This project abstracts these pose-estimation errors into controlled reward corruptions. Gaussian jitter represents frame-to-frame instability, dropout represents missing detections, and systematic bias represents camera or estimator offsets.

Pose-driven humanoid control under noisy inputs. Recent humanoid-control work has begun to address the difficulty of controlling simulated bodies from imperfect pose inputs. Perpetual Humanoid Control is designed for real-time avatar control from video or language-derived poses and discusses noisy pose estimates, occlusion, challenging viewpoints, fast motion, and physically implausible poses as major challenges for humanoid control (Luo et al., 2023). This line of work is closely related to the motivation of using pose signals for embodied control, but it focuses on building robust humanoid controllers. In contrast, this project isolates the reward side of the problem by keeping observations clean and corrupting only the keypoints used to compute the dense imitation reward.

Reward corruption and reinforcement learning. Reward corruption has been studied in reinforcement learning as a source of misspecification, robustness failure, and unsafe optimization. Everitt et al. formalize corrupted reward channels and show that observed rewards may differ systematically from the intended reward (Everitt et al., 2017). Wang et al. study reinforcement learning with perturbed rewards and propose methods for learning when the observed reward channel is noisy or biased (Wang et al., 2020). Pose-derived reward corruption has a specific structure because the reward is computed from corrupted intermediate keypoint measurements rather than perturbed after computation. This means that the rule used to handle missing or biased keypoints can change the

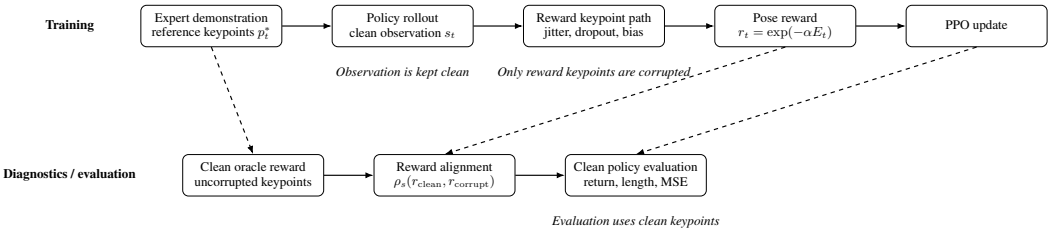

Figure 1: Experimental design. The policy observes the clean simulator state, while only the keypoints used to compute the scalar pose reward are corrupted. Reward alignment compares the corrupted training reward with the clean oracle reward, and final policy evaluation is performed under clean keypoints.

reward function itself. The present project focuses on this gap by testing whether corrupted pose rewards preserve the same behavioral ordering as the clean oracle reward.

PPO and demonstration-guided control. Proximal Policy Optimization is a widely used on-policy policy-gradient algorithm for continuous control (Schulman et al., 2017). I use PPO as the downstream optimizer because it directly optimizes the scalar reward produced by the pose-reward wrapper, making it a useful test case for corrupted pose rewards.

3 Method

3.1 Problem Setup

The project considers demonstration-guided reinforcement learning in continuous-control environments. Let s_t denote the simulator state observed by the policy, a_t the action, and p_t the set of simulator body keypoints extracted at time t. An expert demonstration provides a reference keypoint trajectory p_t^* for $t = 1, \ldots, T$. PPO trains the policy using a dense pose reward that compares the policy keypoints to the expert keypoints.

The key experimental control is that the policy observation remains clean. Corruption is applied only to the keypoints used in the reward computation. This design isolates reward-side corruption, so differences across conditions can be attributed to the scalar reward signal rather than noisy observations.

3.2 Root-Relative Keypoint Representation

For each environment, I extract selected MuJoCo body-center positions from `data.xpos` and convert them to root-relative planar keypoints. If $x_{t,k}$ is the world-frame position of body keypoint k at time t and $x_{t,\text{root}}$ is the root body position, the root-relative keypoint is

$$p_{t,k} = \Pi(x_{t,k} - x_{t,\text{root}}), \tag{1}$$

where Π keeps the MuJoCo x-z coordinates. MuJoCo body positions are represented as $[x, y, z]$, and the implementation keeps horizontal position and vertical height. The root body is the first selected body, `torso`, for all three environments. This representation removes global translation and focuses the reward on body configuration relative to the torso.

Walker2d-v5 uses seven torso, leg, and foot body centers. Hopper-v5 uses four torso-to-foot body centers. Humanoid-v5 uses thirteen torso, pelvis, leg, foot, and arm body centers. The Humanoid keypoint set supports arm-specific systematic bias, which connects the experiment to full-body sports feedback where upper-limb tracking errors matter.

3.3 Pose-Derived Dense Reward

The clean pose reward compares the policy keypoints to the expert reference keypoints at the corresponding trajectory phase. At environment step t, the wrapper selects the reference frame

$$t_{\text{ref}} = \min(t, T - 1),$$

where T is the reference trajectory length. In all training and evaluation calls, `terminate_on_reference_end` is set to `True`, so episodes terminate once the reference horizon is reached unless the MuJoCo environment terminates earlier. The reference trajectory is not looped.

The pose error is

$$E_t = \sum_{k=1}^{K} w_k \left\| p_{t,k} - p^*_{t_{\mathrm{ref}},k} \right\|_2^2, \tag{2}$$

and the reward is

$$r_t = \exp(-\alpha E_t). \tag{3}$$

I use $\alpha = 5.0$ and uniform keypoint weighting with $w_k = 1$ for all selected body keypoints. During evaluation, the clean pose reward is computed using uncorrupted simulator keypoints.

3.4 Reward Corruption Models

I evaluate structured keypoint corruption models that correspond to common pose-estimation errors in sports-video settings.

Gaussian jitter. Gaussian noise models frame-to-frame keypoint instability from fast motion:

$$\tilde{p}_{t,k} = p_{t,k} + \epsilon_{t,k}, \qquad \epsilon_{t,k} \sim \mathcal{N}(0, \sigma^2 I). \tag{4}$$

Dropout-zero. Dropout-zero models a poor missing-keypoint handling rule. Each keypoint is dropped with probability p_{drop}, and missing keypoints are replaced with zero before computing the reward. This introduces artificial pose errors because zero is an arbitrary coordinate rather than an estimate of the true body location.

Dropout-hold. Dropout-hold models an optimistic tracking-memory baseline. Missing keypoints are replaced with the previous visible keypoint value. This preserves temporal continuity and avoids origin-based artifacts, but it may be more optimistic than a real pose-estimation pipeline.

Dropout-masked. Dropout-masked excludes missing keypoints from the pose error. With visibility mask $m_{t,k}$, the masked pose error is

$$E_t^{\mathrm{masked}} = \frac{\sum_{k=1}^{K} m_{t,k} w_k \left\| \tilde{p}_{t,k} - p^*_{t,k} \right\|_2^2}{\sum_{k=1}^{K} m_{t,k} w_k + \epsilon}. \tag{5}$$

The reward is then $\exp(-\alpha E_t^{\mathrm{masked}})$. This is equivalent to a binary confidence-weighted reward and naturally extends to continuous keypoint confidence scores.

Systematic bias. Systematic bias models viewpoint-dependent or estimator-dependent offsets:

$$\tilde{p}_{t,k} = p_{t,k} + b_k. \tag{6}$$

In Walker2d and Hopper, this is applied to foot keypoints. In Humanoid, I also test arm-specific bias.

3.5 Reward-Task Alignment

To measure whether corrupted rewards preserve the intended behavior ordering, I compute reward alignment between the clean oracle reward and the corrupted training reward over matched rollout states. For a given environment, reference trajectory, policy, and corruption condition, I roll out the loaded PPO policy deterministically for five episodes. At each visited timestep t, I extract the clean simulator keypoints, select the corresponding reference keypoints p_t^*, and compute both the clean pose reward and the corrupted reward on that same state. The primary metric is Spearman rank correlation:

$$\rho_s = \mathrm{Spearman}(r_{\mathrm{clean}}, r_{\mathrm{corrupt}}). \tag{7}$$

High Spearman alignment means that the corrupted reward ranks states similarly to the clean reward. Low Spearman alignment means that corruption changes the ordering of behavior optimized by PPO.

Table 1: Expert policies and deterministic reference trajectories used in the experiments. Reference return is the original MuJoCo environment return accumulated by the saved demonstration trajectory.

Environment	Expert source	Expert budget	Reference length	Reference return
Walker2d-v5 early	PPO	10k steps	159	278.8
Walker2d-v5 main	PPO	1M steps	1000	3469.1
Hopper-v5	PPO	500k steps	1000	3421.3
Humanoid-v5	RL Zoo PPO + VecNormalize	5M steps	1000	6639.5

I use this diagnostic in two ways. For the dropout-rate sweep, I use fixed expert-policy rollouts for each environment: the 1M-step Walker2d expert, the 500k-step Hopper expert, and the RL Zoo Humanoid expert. I then recompute alignment for each dropout rate, missing-keypoint rule, and random seed using the same environment-specific reference trajectory. This keeps the behavior distribution fixed while comparing how zero-imputation, masked scoring, and hold-last imputation change the reward ordering. For the alignment-performance analysis, I use each trained condition's own policy and join its alignment score with its clean-evaluation performance. This tests whether reward misalignment under the policy's own state distribution corresponds to downstream PPO behavior.

I also compute Pearson correlation and mean corrupted reward for diagnostics, but the main analysis focuses on Spearman correlation because reward scale can change without preserving reward ordering.

4 Experimental Setup

4.1 Environments

I evaluate the pose-reward pipeline in three MuJoCo locomotion environments.

Walker2d-v5. Walker2d is the main controlled environment. Its clean pose reward is learnable, but the environment remains sensitive to reward corruption. I use Walker2d to test whether low reward alignment corresponds to downstream PPO errors and whether masked scoring mitigates zero-imputed dropout.

Hopper-v5. Hopper is used as a second morphology. It has fewer keypoints and a stable 1000-step expert reference. Hopper tests whether poor reward alignment necessarily implies downstream failure or whether simpler morphologies can tolerate reward distortion.

Humanoid-v5. Humanoid is used as a full-body stress test. It includes torso, pelvis, leg, foot, and arm keypoints, making it more relevant to full-body athletic pose feedback. A stable Humanoid expert was obtained using RL Baselines3 Zoo PPO with observation normalization. Pose-reward PPO imitation remains difficult in this high-dimensional morphology, so Humanoid is interpreted as a stress test rather than the cleanest causal environment.

4.2 Expert Demonstrations

For each environment, I train or load a default-reward PPO expert and collect a deterministic reference trajectory. The initial Walker2d experiment used a short 159-step reference from an early 10k-step PPO expert. I treat this as first-pass validation that the pose-reward wrapper, keypoint extraction, reference collection, and PPO training loop worked end-to-end. The main Walker2d experiments use a stronger 1000-step reference collected from a 1M-step default-reward PPO expert. Hopper uses a 1000-step reference from a stable 500k-step PPO expert. Humanoid uses a 1000-step reference from an RL Baselines3 Zoo PPO expert trained with observation normalization.

4.3 Training and Evaluation

All pose-reward policies are trained with Stable-Baselines3 PPO using a single `DummyVecEnv` wrapped with `VecMonitor`. During training, the original MuJoCo reward is replaced by the pose-derived reward, while the original environment reward is preserved in the `info` dictionary for

Table 2: Spearman alignment between clean and corrupted pose rewards across dropout rates. Values are averaged over five random seeds.

Environment	Rule	$p = 0.1$	$p = 0.2$	$p = 0.3$	$p = 0.5$	$p = 0.7$
Hopper	Hold	0.975	0.949	0.922	0.861	0.800
Hopper	Masked	0.928	0.837	0.739	0.488	0.239
Hopper	Zero	0.540	0.285	0.147	0.045	0.027
Humanoid	Hold	1.000	1.000	1.000	0.999	0.999
Humanoid	Masked	0.937	0.887	0.840	0.724	0.559
Humanoid	Zero	0.586	0.375	0.282	0.130	0.026
Walker2d	Hold	0.999	0.998	0.997	0.994	0.990
Walker2d	Masked	0.983	0.959	0.927	0.807	0.563
Walker2d	Zero	0.496	0.287	0.197	0.126	0.107

evaluation. The PPO policy is `MlpPolicy` with the default Stable-Baselines3 MLP architecture. I use rollout length 1024, batch size 64, 10 optimization epochs per rollout, learning rate 3×10^{-4}, discount factor $\gamma = 0.99$, GAE parameter $\lambda = 0.95$, clipping range 0.2, and entropy coefficient 0.0.

Each condition is trained across three random seeds. Most pose-reward policies are trained for 200k environment steps. The stronger-reference Walker2d follow-up additionally includes 500k-step runs to test whether the longer 1000-step reference benefits from a larger optimization budget. Final evaluation is always clean: corrupted-reward policies are evaluated with uncorrupted simulator keypoints and the clean pose reward. Policies are evaluated deterministically for five episodes using `model.predict(..., deterministic=True)`. The primary metrics are clean pose return, original MuJoCo environment return, episode length, keypoint MSE against the expert reference, and Spearman reward alignment.

Most Walker2d and Hopper pose-reward experiments were run with the local single-environment PPO pipeline. The longer Humanoid expert training used RL Baselines3 Zoo with observation normalization and was launched through Modal in an L4-backed container.

5 Results

5.1 Result 1: Missing-Keypoint Handling Determines Reward Alignment

The first result tests whether dropout harms the reward because keypoints are missing or because missing keypoints are handled poorly. I sweep the dropout rate $p_{\text{drop}} \in \{0.1, 0.2, 0.3, 0.5, 0.7\}$ and compare zero-imputation, masked scoring, and hold-last imputation across Walker2d, Hopper, and Humanoid.

The results show a consistent hierarchy: $\text{hold} > \text{masked} > \text{zero}$. Zero-imputation rapidly collapses reward alignment as dropout increases. Even at low dropout rates, replacing missing keypoints with zeros substantially reduces Spearman correlation between clean and corrupted rewards. By $p = 0.30$, zero-imputation has low alignment in all three morphologies. Masked scoring preserves more alignment at mild and moderate dropout rates, while hold-last imputation preserves the most alignment.

This result supports the main reward-design claim. Missing keypoints are not harmful only because information is absent. They become especially harmful when the reward assigns arbitrary coordinates to missing values, causing PPO to optimize a reward that can rank behavior differently from the clean imitation objective. Figure 2 visualizes this pattern in Walker2d-v5, and Table 2 reports the same sweep across all three environments.

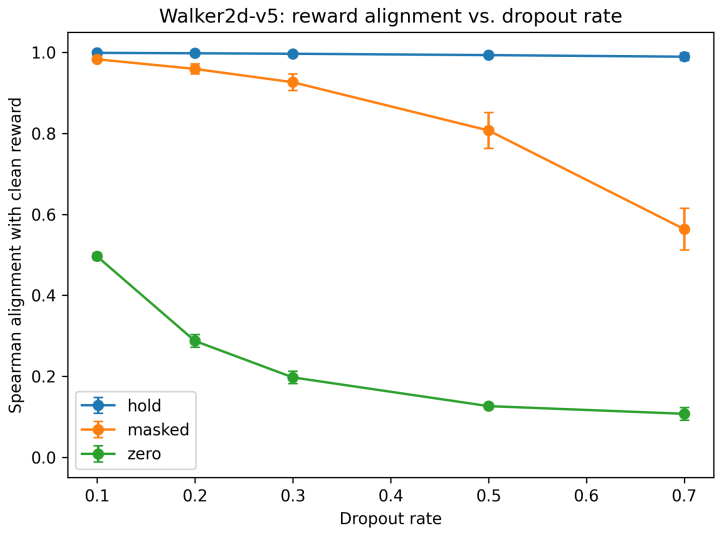

Figure 2: Walker2d-v5 dropout-rate sweep for reward alignment. Zero-imputation rapidly collapses Spearman alignment as missingness increases, while masked scoring degrades more gradually and hold-last imputation preserves the most reward structure. Table 2 shows that the same hold $>$ masked $>$ zero hierarchy appears across Hopper-v5, Humanoid-v5, and Walker2d-v5.

5.2 Result 2: Masked Scoring Mitigates Zero-Imputed Dropout in Walker2d

Walker2d provides the clearest test of whether low reward alignment corresponds to downstream PPO errors because the clean pose reward is learnable while still sensitive to corruption. I use the stronger 1000-step Walker2d reference and compare clean reward training with two missing-keypoint treatments at the same dropout rate, $p = 0.30$.

At the 500k-step training budget, masked dropout improves average clean-evaluation environment return from 238.8 under zero-imputation to 915.6 and reduces keypoint MSE from 0.1257 to 0.0582. This supports the alignment story: zero-imputation creates misleading reward structure, while masking missing keypoints avoids penalizing the agent for arbitrary origin-based artifacts. Figure 5 provides a qualitative trajectory-level comparison of the learned Walker2d foot motion under zero-imputed and masked dropout. However, the masked condition has high variance across seeds, with pose return 138.9 ± 112.0 and environment return 915.6 ± 470.4. The result should therefore be interpreted as evidence of a mitigation trend rather than a statistically conclusive improvement. More seeds or longer training would be needed to make a stronger claim about final performance.

Table 3: Walker2d mitigation comparison at $p = 0.30$ using the stronger 1000-step reference and a 500k-step PPO training budget. Values are averaged over three seeds.

Condition	Pose Return	Env. Return	Keypoint MSE	Episode Length
Clean	59.5 ± 4.5	448.6 ± 59.3	0.0661 ± 0.0277	299.5 ± 77.2
Dropout-zero	48.5 ± 22.2	238.8 ± 98.1	0.1257 ± 0.0936	265.3 ± 96.1
Dropout-masked	$\mathbf{138.9 \pm 112.0}$	$\mathbf{915.6 \pm 470.4}$	$\mathbf{0.0582 \pm 0.0113}$	$\mathbf{614.3 \pm 292.3}$

5.3 Result 3: Cross-Environment Results Show Alignment Is Diagnostic but Not Sufficient

The final analysis asks when low reward alignment predicts downstream PPO errors. The answer depends on whether reward corruption is the main bottleneck. In Walker2d, the clean pose reward is learnable and the environment remains sensitive to reward corruption, making it the clearest setting for the alignment-performance claim. Zero-imputed dropout lowers reward alignment and worsens clean-evaluation keypoint MSE, while masked scoring preserves more reward structure and improves final performance.

Table 4: Reward alignment across environments at representative corruption levels. These Spearman values are computed on each trained condition's own deterministic rollout distribution and are used for the alignment-performance analysis.

Condition	Walker2d	Hopper	Humanoid
Clean	1.000	1.000	1.000
Jitter $\sigma = 0.10$	0.800	0.631	0.762
Hold dropout $p = 0.30$	0.999	1.000	0.999
Zero dropout $p = 0.30$	**0.359**	**0.232**	**0.230**
Masked dropout $p = 0.30$	0.891	0.648	0.905
Foot bias 0.10	0.967	0.562	0.983

Hopper qualifies this pattern. Dropout-zero at $p = 0.30$ has low Spearman alignment, but the trained policies still achieve stable 1000-step behavior and low clean-evaluation keypoint MSE. The clean Hopper baseline is also highly variable, with environment return 702.6 ± 360.7 and episode length 578.1 ± 384.0, suggesting that some seeds learn the reference while others fall into brittle local optima. One possible explanation is that the clean pose reward is sharper and less forgiving near the reference trajectory, while some corrupted rewards smooth or reshape the objective enough to improve optimization in this simpler morphology.

Humanoid shows the opposite limitation. The clean pose-reward condition has perfect alignment by definition, but pose-reward PPO remains short-horizon in the high-dimensional full-body setting. Zero-imputed dropout again produces the clearest reward pathology, reducing Spearman alignment to approximately 0.230, while masked dropout preserves much higher alignment. However, masked dropout does not improve final keypoint MSE relative to clean in Humanoid. This shows that preserving reward ordering is not sufficient to solve high-dimensional full-body imitation.

Together, these results show that low alignment is most predictive when the clean pose-reward task is learnable and corrupted rewards are the main source of failure. Low alignment can flag a misleading reward, but high alignment does not guarantee PPO success. Table 4 summarizes the representative alignment pattern. Figure 3 in the appendix visualizes this alignment pattern across environments, and Figure 4 shows the corresponding clean-evaluation keypoint MSE. The full Hopper and Humanoid performance matrices are also included in the appendix.

6 Discussion

The experiments support three main conclusions. First, missing-keypoint handling can change the reward ordering optimized by PPO. Zero-imputation is especially harmful because it replaces missing keypoints with arbitrary coordinates, creating artificial pose errors unrelated to the true deviation from the expert. Masked scoring avoids this failure by only scoring visible keypoints, while hold-last imputation avoids origin-based artifacts by preserving temporal continuity. Hold-last should be interpreted as an optimistic tracking-memory baseline rather than a realistic pose-estimator default.

Second, reward alignment is useful but incomplete. In Walker2d, low Spearman alignment is a meaningful warning sign because the corrupted reward no longer matches the clean imitation objective and PPO performance degrades. Hopper shows that low alignment can be tolerated in simpler settings, while Humanoid shows that high alignment does not guarantee successful imitation. Reward alignment is therefore best understood as a diagnostic for reward reliability, not as a full predictor of final control performance.

Third, morphology and optimization difficulty matter. Hopper remains robust even under reward distortion, while Humanoid remains difficult even with clean rewards. This suggests that dense pose matching alone may be insufficient for full-body imitation without stronger learning machinery such as behavior cloning initialization, phase alignment, curriculum learning, hybrid task rewards, or adversarial motion priors.

6.1 Limitations

This study isolates reward-side corruption by keeping policy observations clean. This makes the experimental comparison cleaner, but real pose-based systems may have both noisy observations and noisy rewards. The corruption models are synthetic and approximate jitter, dropout, and bias, but they are not learned from real sports-video pose tracks. Cross-environment comparisons should be interpreted qualitatively because Walker2d, Hopper, and Humanoid differ in morphology, reference trajectory quality, and optimization difficulty.

A further limitation is the use of fixed frame-by-frame reference indexing. The reward compares the policy at step t to the expert reference frame at step t, and the reference is not phase-aligned or time-warped. If a policy falls behind or moves ahead of the expert, part of the reward error may reflect temporal phase mismatch rather than only pose mismatch. This is especially relevant for longer 1000-step references and for comparing conditions with different episode lengths. Finally, Humanoid pose-reward PPO remains difficult even with a strong expert reference, which limits conclusions about full-body policy quality.

6.2 Future Work

Future work should learn realistic keypoint error models from real sports-video pose estimators and test whether the same reward-alignment failures occur on real pose tracks. A natural next step is to replace binary masking with confidence-weighted scoring, where each keypoint contributes in proportion to pose-estimator confidence. Future experiments should also corrupt both policy observations and reward keypoints. For full-body settings, stronger imitation-learning machinery may be needed, including behavior cloning warm starts, phase alignment, temporal warping, curriculum learning, or hybrid rewards that combine pose matching with task-level progress.

7 Conclusion

This project studied how structured pose-estimation errors affect dense pose-derived rewards and PPO policies trained from them. Motivated by pose-based coaching feedback, the experiments isolate reward corruption by corrupting only the reward keypoints while keeping policy observations clean. The main finding is that missing-keypoint handling matters: zero-imputation consistently damages reward alignment, while masked and hold-last scoring preserve more reward structure. In Walker2d, this misalignment corresponds to downstream PPO degradation and masked scoring improves clean-evaluation performance, although the mitigation result has high variance across seeds. Hopper and Humanoid qualify this relationship, showing that reward alignment is diagnostic but not sufficient. Robust pose-derived reinforcement learning requires both reliable reward construction and learning algorithms capable of optimizing the resulting imitation objective.

8 Team Contributions

This project was completed individually by Joseph Dehoney. The author was responsible for problem formulation, related work review, environment setup, expert policy training, keypoint extraction, reward-wrapper implementation, corruption-model implementation, PPO training runs, evaluation scripts, reward-alignment analysis, figure generation, poster preparation, and report writing.

Changes from Proposal. The original proposal focused on studying pose-derived reward corruption in a controlled locomotion setting, with possible extensions to additional environments and robust reward variants. During the project, the scope expanded from the initial Walker2d pipeline to include Hopper-v5 as a second morphology and Humanoid-v5 as a full-body stress test. The final project also placed more emphasis on reward-alignment diagnostics and missing-keypoint handling rules because the experiments showed that zero-imputation, masked scoring, and hold-last imputation produce qualitatively different reward-ordering behavior.

9 AI Tools Disclosure

Generative AI tools, namely ChatGPT, were used as support tools during the project for debugging assistance, LaTeX formatting, and improving clarity and design for tables. The author independently designed the research question, implemented and ran the experiments, generated the results, inspected the outputs, and verified the technical claims. AI tools were not used to fabricate results, labels, experiments, or quantitative findings.

References

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 1 (2019), 172–186. doi:10.1109/TPAMI.2019.2929257

Tom Everitt, Victoria Krakovna, Laurent Orseau, and Marcus Hutter. 2017. Reinforcement Learning with a Corrupted Reward Channel. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. 4705–4713. doi:10.24963/ijcai.2017/656

Zhengyi Luo, Jinkun Cao, Kris Kitani, Weipeng Xu, Hongyi Liu, Shijie Zhou, Xue Bin Peng, Hugo Bérard, and C. Karen Liu. 2023. Perpetual Humanoid Control for Real-time Simulated Avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10895–10904.

Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. 2018. DeepMimic: Example-Guided Deep Reinforcement Learning of Physics-Based Character Skills. In *ACM SIGGRAPH 2018 Papers*. 1–14. doi:10.1145/3197517.3201311

Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. 2021. AMP: Adversarial Motion Priors for Stylized Physics-Based Character Control. *ACM Transactions on Graphics* 40, 4 (2021), 1–20. doi:10.1145/3450626.3459670

France Rose, Monika Michaluk, Timon Blindauer, Bogna M. Ignatowska-Jankowska, Liam O'Shaughnessy, Greg J. Stephens, Talmo D. Pereira, Marylka Y. Uusisaari, and Katarzyna Bozek. 2026. Deep Imputation for Skeleton data (DISK) for behavioral science. *Nature Methods* 23 (2026), 236–247. doi:10.1038/s41592-025-02893-y

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347* (2017).

Jingkang Wang, Yang Liu, and Bo Li. 2020. Reinforcement Learning with Perturbed Rewards. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 6202–6209. doi:10.1609/aaai.v34i04.6086

A Additional Experimental Results

This appendix reports additional cross-environment visualizations and the full Hopper-v5 and Humanoid-v5 corruption matrices. The main paper uses these results to support the conclusion that reward alignment is diagnostic but not sufficient.

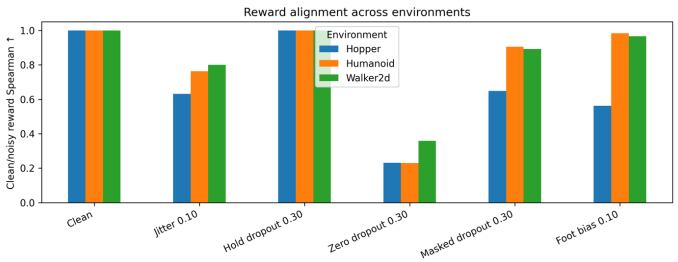

Figure 3: Cross-environment Spearman reward alignment by condition. Zero-imputed dropout produces low reward alignment across morphologies, while masked and hold-last dropout preserve more reward structure.

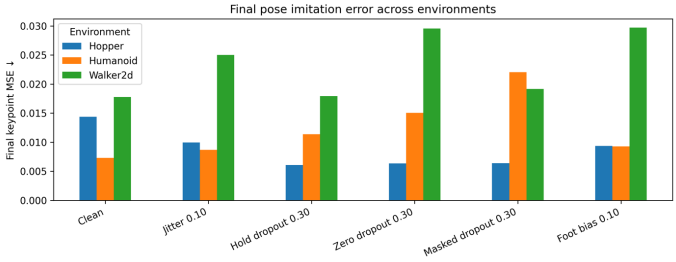

Figure 4: Cross-environment clean-evaluation keypoint MSE by condition. Reward alignment and downstream pose quality do not have the same relationship in Walker2d-v5, Hopper-v5, and Humanoid-v5.

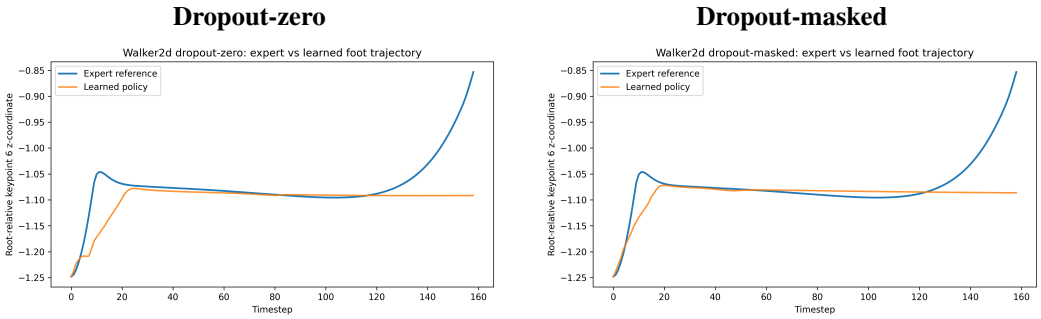

Figure 5: Walker2d-v5 foot-keypoint trajectories under zero-imputed and masked dropout, evaluated with clean simulator keypoints. The comparison illustrates the trajectory-level effect of the missing-keypoint handling rule: zero-imputation can push learning toward distorted motion because missing keypoints are treated as artificial origin targets, while masked scoring avoids assigning a coordinate value to missing detections.

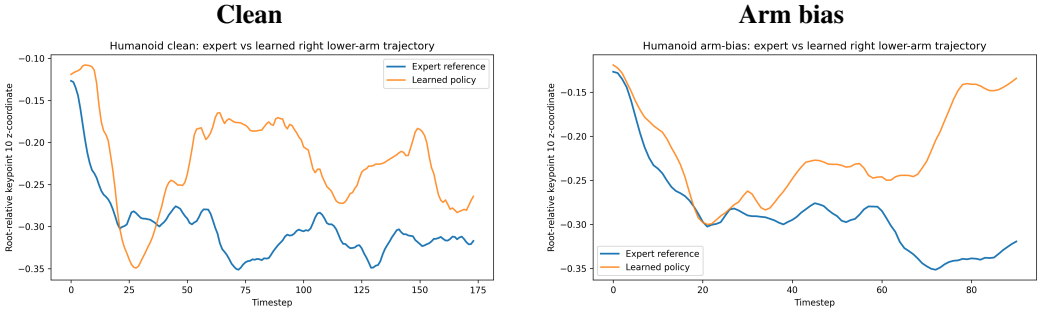

Figure 6: Humanoid-v5 right-lower-arm keypoint trajectories under clean reward training and arm-biased reward training, evaluated with clean simulator keypoints. The comparison illustrates how a systematic upper-limb reward bias can alter the learned arm trajectory even when reward-rank alignment remains high, connecting the Humanoid stress test to sports-form feedback where arm-tracking errors are practically important.

Table 5: Full Hopper-v5 corruption matrix. Policies were trained for 200k PPO steps using a 1000-step reference trajectory collected from a 500k-step default-reward PPO expert and evaluated under the clean pose reward. Values report mean ± standard deviation over three seeds.

Condition	Pose Return	Env. Return	Keypoint MSE	Episode Length	Spearman
Dropout-hold $p = 0.30$	812.31 ± 6.51	1039.04 ± 0.40	$\mathbf{0.0061 \pm 0.0004}$	1000.0 ± 0.0	1.000 ± 0.000
Foot bias 0.05	808.60 ± 1.11	1039.93 ± 1.98	0.0061 ± 0.0002	1000.0 ± 0.0	0.835 ± 0.011
Dropout-zero $p = 0.30$	804.00 ± 11.91	1044.05 ± 9.43	0.0064 ± 0.0004	1000.0 ± 0.0	0.232 ± 0.034
Dropout-masked $p = 0.30$	804.32 ± 4.73	1034.15 ± 1.96	0.0064 ± 0.0001	1000.0 ± 0.0	0.648 ± 0.017
Dropout-hold $p = 0.10$	796.76 ± 16.18	1043.75 ± 4.74	0.0067 ± 0.0006	1000.0 ± 0.0	1.000 ± 0.000
Jitter $\sigma = 0.05$	600.95 ± 364.54	830.70 ± 359.09	0.0088 ± 0.0049	733.9 ± 383.8	0.866 ± 0.059
Foot bias 0.10	600.56 ± 344.46	1249.06 ± 919.01	0.0094 ± 0.0051	712.3 ± 396.2	0.562 ± 0.189
Jitter $\sigma = 0.10$	359.56 ± 374.37	558.80 ± 422.34	0.0100 ± 0.0045	475.9 ± 398.0	0.631 ± 0.151
Clean	435.76 ± 330.26	702.60 ± 360.68	0.0144 ± 0.0078	578.1 ± 384.0	1.000 ± 0.000

Table 6: Full Humanoid-v5 corruption matrix. Policies were trained for 200k PPO steps using a 1000-step reference trajectory collected from an RL Zoo PPO expert with observation normalization and evaluated under the clean pose reward. Values report mean ± standard deviation over three seeds.

Condition	Pose Return	Env. Return	Keypoint MSE	Episode Length	Spearman
Clean	39.23 ± 5.40	350.03 ± 78.77	$\mathbf{0.0073 \pm 0.0023}$	71.9 ± 16.0	1.000
Jitter $\sigma = 0.10$	32.59 ± 7.03	295.84 ± 64.50	0.0087 ± 0.0011	65.0 ± 10.8	0.762
Foot bias 0.10	33.61 ± 8.07	332.30 ± 42.52	0.0093 ± 0.0062	71.1 ± 8.9	0.983
Arm bias 0.10	34.15 ± 5.24	340.89 ± 26.40	0.0109 ± 0.0046	70.5 ± 4.5	1.000
Dropout-hold $p = 0.30$	30.97 ± 8.47	319.23 ± 67.89	0.0114 ± 0.0042	66.8 ± 16.2	0.999
Dropout-hold $p = 0.10$	39.14 ± 13.79	$\mathbf{419.37 \pm 102.08}$	0.0118 ± 0.0047	$\mathbf{85.6 \pm 23.8}$	1.000
Foot bias 0.05	31.84 ± 11.32	355.56 ± 63.29	0.0121 ± 0.0050	74.3 ± 14.2	0.997
Jitter $\sigma = 0.05$	32.85 ± 0.82	351.36 ± 71.04	0.0126 ± 0.0044	72.7 ± 13.4	0.959
Arm bias 0.05	33.35 ± 6.54	355.70 ± 22.75	0.0133 ± 0.0054	72.8 ± 5.1	1.000
Dropout-zero $p = 0.30$	23.95 ± 3.82	309.94 ± 97.70	0.0151 ± 0.0099	66.6 ± 19.9	0.230
Dropout-masked $p = 0.30$	20.63 ± 8.40	361.75 ± 33.46	0.0220 ± 0.0062	75.5 ± 5.3	0.905