DYNAMIC HUMANOID FALL RECOVERY: A DATASET AND ROBUST RECOVERY POLICY

Anonymous authors
Paper under double-blind review

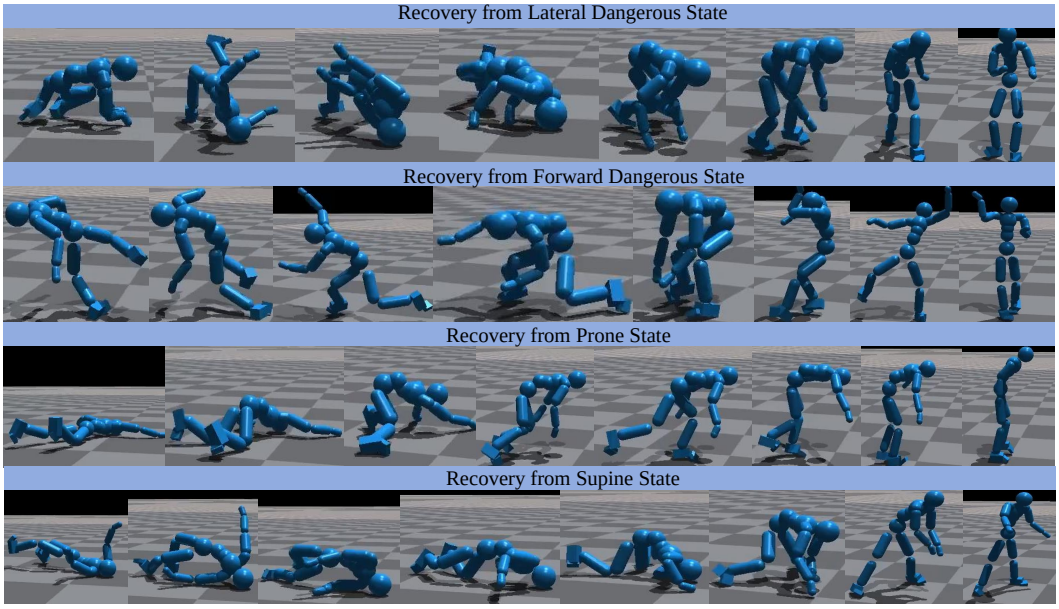


Figure 1: **Humanoid fall recovery capabilities from dynamic states.** While existing methods typically focus on getting up from fixed canonical poses (prone or supine, bottom), the proposed KiPAR generalizes to highly unstable configurations (top) encountered during realistic perturbations.

ABSTRACT

Humanoid fall recovery remains challenging due to the highly diverse and out-of-distribution recovery states from real-world falls, and the conflict between rapid physical recovery and stable, natural motion. Existing approaches mainly address this challenge through multi-stage curriculum learning. However, such pipelines require complex multi-stage configurations and suffer from policy degradation due to distribution shifts induced by rigid stage transitions. To address these limitations, we propose **KiPAR**, a recovery-oriented reinforcement learning framework that combines dynamic state initialization, history-consistent motion priors, and phase-aware reward scheduling. KiPAR bypasses canonical static initializations by using a comprehensive dataset of diverse, dynamic dangerous states as initialization source. To ensure coherence after resetting to these highly unstable states, we introduce kinematic prior history injection to reconstruct physically consistent observation histories. Furthermore, we introduce a phase-aware reward scheme that dynamically balances heuristic incentives and stabilization regularizers. Extensive experiments demonstrate that KiPAR achieves a 98.34% recovery success rate, a 2.77s average time-to-stand, and a low 3.12% recollapse rate.

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.

1 INTRODUCTION

Humanoid robots possess immense potential across a diverse range of applications, spanning industrial automation, healthcare assistance, and hazardous rescue operations (Cao, 2024; Gaspard et al., 2025; Kanehiro et al., 2003; Sheng et al., 2025; Gu et al., 2026). Recent advances in reinforcement learning and motion imitation have enabled humanoids to acquire agile locomotion and complex whole-body behaviors (Peng et al., 2018; 2021; Luo et al., 2023; Wu et al., 2023; Margolis et al., 2024; Luo et al., 2025). However, encountering critical physical instabilities or external disturbances is inevitable when operating in complex and dynamic environments. Equipping humanoids with the capability for rapid and stable recovery is significant to achieve full autonomy.

Despite recent learning-based frameworks (Huang et al., 2025; He et al., 2025b; Chen et al., 2025; Yang et al., 2023; Tran et al., 2018; Kim et al., 2019) having achieved certain accomplishments in recovery, two critical gaps still hinder their deployment in the real world. First, canonical resting-pose initialization does not fully represent disturbed states with substantial linear and angular velocity. Offline snapshots provide access to such states, but a snapshot-only reset does not supply the preceding observations needed by a history-conditioned controller. Second, fall recovery requires balancing task completion with motion regularization. Heuristic rewards can encourage effective but abrupt actions, while an Adversarial Motion Prior (Peng et al., 2021; Ho & Ermon, 2016) trained on a mismatched motion distribution can assign low rewards to necessary recovery movements. Multi-stage pipelines offer one solution (He et al., 2025b), but this architecture requires tedious multi-environment configurations, and abrupt stage transitions invite policy degradation due to distribution shift (Ablett, 2025; Jiang et al., 2023).

These limitations highlight the need for a more realistic benchmark for humanoid fall recovery. Recent studies demonstrate recovery from diverse fallen configurations and terrains (He et al., 2025b; Huang et al., 2025; Xu et al., 2025). However, existing recovery settings do not explicitly characterize the dynamically evolving dangerous states that precede recovery, particularly the coupled variations in body configuration, linear and angular velocities, contact conditions, and support-limb interactions. Such dynamic factors can critically affect the subsequent recovery process, yet they are difficult to capture through manually designed fallen poses or isolated disturbed states.

To address these challenges, we propose KiPAR, an reinforcement learning framework that elegantly reconciles functional recovery with kinematic naturalness. KiPAR uses disturbed states collected from motion-tracking rollouts, Kinematic Prior History Injection for approximate temporal initialization, and a phase-aware reward scheme combining recovery incentives, reference tracking, and stabilization regularizers. The proposed framework eliminates rigid stage transitions and effectively circumvents distribution shifts.

Contributions: 1) Realistic Motion-Driven Fall Dataset: We introduce a highly diverse humanoid failure dataset generated by destabilizing pretrained motion tracking policy under dynamic external perturbations. Unlike conventional static or synthetically perturbed initializations, our dataset comprehensively captures the physically plausible, dynamic manifolds of real-world falls, providing a robust foundation for unconstrained recovery training. **2) Kinematic Prior History Injection:** To resolve the fundamental kinematic contradictions caused by environmental resets, we propose Kinematic Prior History Injection, a backward kinematic extrapolation to initialize a history-conditioned policy from a dynamic snapshot. **3) Phase-Aware Reward and Stage-Gated Regularization:** We formulate a framework that dynamically balances recovery robustness, motion-prior regularization, and stabilization objectives throughout the recovery process.

2 RELATED WORK

Humanoid Fall Recovery: Early humanoid fall-recovery systems primarily rely on predefined routines (Kanehiro et al., 2003; Stückler et al., 2006; Fujiwara et al., 2003) or motion planning under kinematic constraints (Kanehiro et al., 2007). Concurrently, optimization-based approaches rely on Zero Moment Point (ZMP) (Pratt et al., 2006) and Model Predictive Control (MPC) to improve disturbance rejection and balance recovery (Katayama et al., 2023; Schwenzer et al., 2021; Romualdi et al., 2022; Al Borno et al., 2012). While effective against moderate perturbations in structured environments, these methods degrade on catastrophic falls due to the difficulty of modeling nonlinear recovery dynamics (Krotkov et al., 2018). Recognizing the fragility of classical con-

trol, the research paradigm has shifted toward Deep Reinforcement Learning, which allows agents to learn robust recovery policies through extensive interactions in highly parallelized physics simulators (e.g., Isaac Gym/Lab (Makoviychuk et al., 2021; Mittal et al., 2025) and MuJoCo (Todorov et al., 2012)). HoST (Huang et al., 2025) learns stand-up motions from scratch by leveraging a multi-critic architecture. HUMANUP (He et al., 2025b) proposes a curriculum-driven, two-stage method to overcome the inherent obstacles of complex contact patterns and sparse reward structures in stand-up tasks. HiFAR (Chen et al., 2025) employs a Multi-Stage Curriculum Learning strategy combined with Key State Initialization to systematically decompose the high-dimensional fall recovery task. Although notable progress has been made in humanoid robot recovery, most studies initialize training from manually specified prone, supine, or kneeling poses (Huang et al., 2025; He et al., 2025b; Chen et al., 2025). Such static initialization simplifies exploration and improves stability, yet it overlooks the temporal and physical continuity between falling and recovery.

Learning-Based Fall Recovery: Research on fall recovery primarily relies on purely heuristic physical exploration, which heavily depends on explicit, hand-crafted reward functions to guide agent behavior (He et al., 2025b; Chen et al., 2025; Gaspard et al., 2025; Wang et al., 2025). This approach not only requires tedious, expert-level reward shaping but also leaves agents prone to exploiting simulator physics loopholes to maximize rewards, resulting in unnatural and mechanically hazardous stand-up strategies. Motion imitation provides behavioral structure beyond task-specific reward shaping. Reference-tracking approaches train physically simulated humanoids to reproduce demonstrated motions (Peng et al., 2018; Luo et al., 2023). AMP (Peng et al., 2021) learns a motion prior that rewards agreement with a reference motion distribution. Under normal locomotion conditions, AMP successfully forces humanoids to execute fluid and natural behaviors (He et al., 2024; 2025a; Luo et al., 2025; Peng et al., 2022). However, general motion datasets predominantly consist of upright, controlled movements (e.g., walking, running) and severely lack unstructured, high-impact fall recovery sequences (Mahmood et al., 2019; Ionescu et al., 2013; Plappert et al., 2016). A motion prior trained on a mismatched reference distribution provides weak incentives for necessary recovery behaviors.

3 Dynamic Fall Recovery Policy

We propose **KiPAR**, an adaptive PPO framework designed for humanoid recovery from diverse, dynamic and severe fallen states. KiPAR aims to reconcile generating sufficient recovery torque with maintaining human-like naturalness. As shown in Figure 2, given a diverse offline dataset of dynamic dangerous states $\mathcal{S}_{fall}$, KiPAR firstly employs an *Adaptive Curriculum Sampler* to strain scenarios based on the current proficiency of the agent. Then a temporal alignment module synchronizes the corresponding reference motion to establish a tracking goal. To address policy oscillation caused by sudden resets to extreme postures, the framework proposes a *Kinematic-Prior History Injection* module (Sec. 3.1) to synthesize a short approximate history by backward extrapolation under the instantaneous velocity field. Finally, a *Phase-Aware Recovery Reward* design (Sec. 3.2) smoothly anneals selected heuristic recovery incentives while retaining a fixed AMP reward weight as the robot approaches stabilization. 69-D Actions over 23 three-DoF joints are mapped to joint-position targets and tracked by a PD controller.

3.1 Kinematic Prior History

Observation Space with Temporal History: Due to the high non-linear of the fall recovery task, a single-frame snapshot is insufficient to capture the full latent dynamics. Therefore, the policy is conditioned on a structured temporal history buffer H_t spanning k consecutive frames:

$$H_t = (o_{t-k+1}, o_{t-k+2}, \ldots, o_t) \in \mathcal{O}^k, \tag{1}$$

where each observation frame o_i contains root height h_{root}, root linear and angular velocities, root orientation vector projection z_{proj} aligned with the global gravity axis, Degrees of Freedom (DoF) positions q, and joint velocities $\dot{q}$.

History Construction: When resetting the humanoid to dynamic unstable snapshots, there is no real preceding trajectory, which causes an abrupt history buffer jump. Conventional paradigms replicate the initial state to populate the historical observation buffer. This approach misleads the

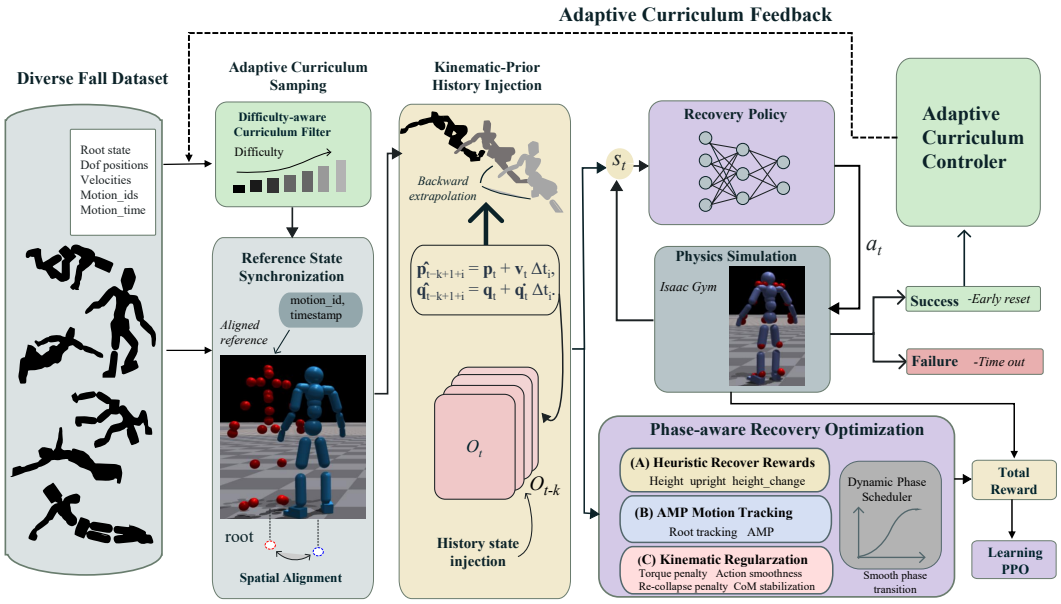


Figure 2: **Overview of the humanoid fall recovery framework.** A diverse dangerous-state dataset is filtered through an adaptive curriculum sampler to progressively select scenarios of ascending difficulty. For each sampled frame, the reference motion is synchronized via temporal alignment and an approximate kinematic observation history is synthesized via history injection. Training combines phase-aware heuristic rewards, reference tracking, AMP reward and gated regularization. Recovery success rate feeds back into the system to adapt the curriculum difficulty distribution.

network into perceiving a completely static history despite the presence of a substantial instantaneous velocity, while physics warm-up rollouts delay policy intervention, risking severe falls and hardware damage. We synthesize a short history window using a kinematic prior computed from the current state and velocities. Let $\mathbf{p}_t, \mathbf{q}_t$ be the current root and joint positions, and $\mathbf{v}_t, \dot{\mathbf{q}}_t$ their velocities. We define k as the total number of frames including the current frame. For $i \in \{0, \ldots, k-1\}$, let $\Delta t_i = (i-k+1)\Delta t$ and approximate

$$\widehat{\mathbf{p}}_{t-k+1+i} = \mathbf{p}_t + \mathbf{v}_t \Delta t_i, \quad \widehat{\mathbf{q}}_{t-k+1+i} = \mathbf{q}_t + \dot{\mathbf{q}}_t \Delta t_i. \tag{2}$$

The $k-1$ preceding states are synthesized, while the last frame is the unmodified current state. Root orientation and velocities are held constant in this approximation.

3.2 Phase-Aware Recovery

Humanoid fall recovery presents an inherent optimization conflict between exploratory recovery behaviors and natural motion regularization. We propose a Phase-Aware Recovery Optimization framework that dynamically transitions the learning objective from physics-driven exploration to motion-prior-constrained stabilization. The objective is to optimize the policy parameters θ to maximize the expected discounted cumulative return, yielding the optimal policy θ^*:

$$\theta^* = \arg\max_{\theta} \mathbb{E}_{\substack{s_0 \sim S_{fall} \\ a_t \sim \pi_\theta, s_{t+1} \sim \mathcal{P}}} \Big[\sum_{t=0}^{T} \gamma^t R_{\text{total},t} \Big]. \tag{3}$$

The training reward combines the environment reward with a separately computed AMP reward:

$$\begin{aligned} R_{\text{total}} &= w_{\text{task}} R_{\text{env}} + w_{\text{disc}} r_t^{\text{amp}}, \\ R_{\text{env}} &= R_{\text{track}} + \beta(\alpha) R_{\text{heuristic}} + w_s R_s - R_{\text{reg}} + R_{\text{success}}, \end{aligned} \tag{4}$$

where $w_{\text{task}} = w_{\text{disc}} = 0.5$, R_{track} denotes the reference-conditioned task reward, $R_{\text{heuristic}}$ represents recovery-oriented heuristic objectives, R_{reg} contains kinematic regularization terms and

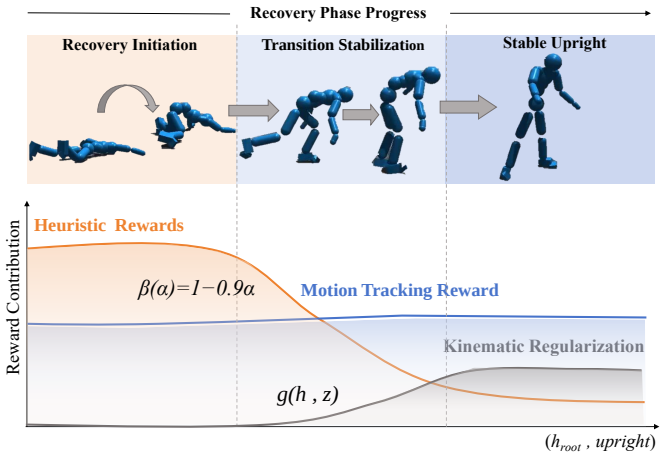
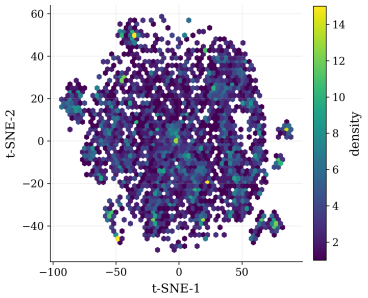
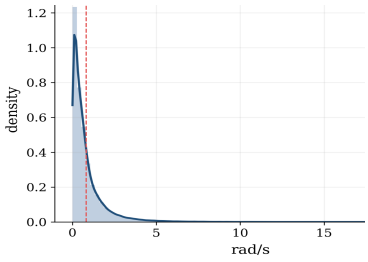


Figure 3: **Phase-aware Reward Scheduling.** Heuristic recovery rewards are emphasized during the early stage and progressively annealed as the humanoid approaches an upright configuration. Motion-tracking and AMP rewards remain active throughout recovery, while kinematic regularization terms are gradually activated through state-dependent gating. This design balances rapid stand-up behavior with natural and physically plausible motion generation.



(a) t-SNE pose distribution



(b) Root angular velocity distribution

Figure 4: Dataset analysis.

R_{success} is the early-success bonus. The phase coefficient $\beta(\alpha)$ modulates the contribution of heuristic rewards based on the current recovery phase α, which is derived from the humanoid's root height:

$$\alpha = \text{clip}\Big(\frac{h_{\text{root}} - h_{\min}}{h_{\max} - h_{\min}}, 0, 1\Big), \tag{5}$$

where h_{root} is the pelvis height, $h_{\min} = 0.3$m corresponds to lying states, and $h_{\max} = 0.85$m corresponds to approximately upright configurations.The phase coefficient controls the relative contribution of heuristic rewards:

$$\beta(\alpha) = 1 - 0.9\alpha. \tag{6}$$

Heuristic objectives dominate optimization during low-height recovery phases, while motion priors gradually take over as the humanoid approaches stable standing. The dynamic evolution of each reward component across different recovery states is illustrated in Figure 3.

AMP Reward: To maintain human-like agility during the recovery, we incorporate an Adversarial Motion Prior (AMP) (Peng et al., 2021). Let ξ_t denote the discriminator's motion-feature history, and let $D(\xi_t) \in (0, 1)$ be its sigmoid probability of belonging to the reference motion distribution. With discriminator-reward normalization disabled, the implemented prior reward is

$$r_t^{\text{amp}} = -s_{\text{disc}} \log(\max\{1 - D(\xi_t), \epsilon\}) . \tag{7}$$

Heuristic Recovery Rewards: The heuristic reward component encourages the policy to rapidly escape fallen configurations and regain balance:

$$R_{\text{heuristic}} = w_h R_h + w_u R_u + w_v R_v, \tag{8}$$

where R_h encourages root elevation, R_u promotes upright torso orientation, and $R_v = 1 - \exp(-2\max(v_z, 0))$ rewards upward root velocity. The survival term $w_s R_s$ is added separately and is not annealed by β.

These objectives provide dense exploration signals during the early recovery phase but are progressively annealed to avoid dominating the final standing behavior.

Motion Tracking Reward: The PHC-based task reward uses point-goal progress and near-reference imitation, with a mechanical-power penalty:

$$R_{\text{track}} = R_{\text{pg}} + R_{\text{power}} + 0.5\,\mathbb{I}[d_t \le 0.25\,\text{m}]R_{\text{imit}}. \tag{9}$$

where d_t is the root-position distance to the synchronized reference, and R_{imit} is the PHC imitation reward (Luo et al., 2023).

Kinematic Regularization: To suppress high-frequency artifacts and physically implausible behaviors, we introduce kinematic regularization terms that are gated by recovery phase.

$$R_{\text{reg}} = R_{\text{action}} + R_{\tau} + R_{\text{rate}} - R_{\text{com}} - R_{\text{stable}} + R_{\text{fall}}, \tag{10}$$

$$g_j(h, z) = \text{clip}\Big(\frac{h - h_j^-}{h_j^+ - h_j^-}, 0, 1\Big) \cdot \text{clip}\Big(\frac{z - z_j^-}{z_j^+ - z_j^-}, 0, 1\Big), \tag{11}$$

where $z = z_{\text{proj}}$ and $j \in \{\text{pen}, \text{com}, \text{stable}\}$ indexes separate gates with term-specific thresholds (Details in Appendix B.3). The action-magnitude and action-rate penalties are

$$R_{\text{action}} = g_{\text{pen}} w_{\text{action}} \|a_t\|^2, \qquad R_{\text{rate}} = g_{\text{pen}} w_{\text{rate}} \|a_t - a_{t-1}\|^2, \tag{12}$$

and the torque penalty is formulated as:

$$R_{\tau} = g_{\text{pen}} w_{\tau} \|\tau\|^2. \tag{13}$$

The support reward uses the mean rigid-body position as a CoM proxy and a heuristic support region constructed from foot contacts:

$$R_{\text{com}} = w_{\text{com}} g_{\text{com}} \max\{0, \tanh\left((r_{\text{sup}} + m - d_{\text{com}})/s\right)\}, \tag{14}$$

where r_{sup} is the maximum horizontal distance of active support bodies from their mean position, d_{com} is the horizontal distance of the CoM proxy to that center, and m, s are margin and scale factors. If no support contact is detected, all candidate support bodies are used. The stable-standing reward suppresses residual momentum:

$$R_{\text{stable}} = w_{\text{stable}} g_{\text{stable}} \exp(-k_{\text{lin}} \min\{\|v\|, c_v\}) \cdot \exp(-k_{\text{ang}} \min\{\|\omega\|, c_\omega\}), \tag{15}$$

where v and ω are root linear and angular velocities, respectively.

4 Experiment

4.1 Dataset Construction

We generate dangerous states by running PHC (Luo et al., 2023) policy with randomized external pushes. In each episode, we sample a push target from a weighted set of body groups (pelvis/torso, head, thighs/calves), and apply a force with randomized magnitude, direction, vertical component, duration, and interval. The simulator runs until PHC triggers termination, and we label a lookback window of frames preceding the termination as dangerous. Each frame is tagged with its motion identity and time stamp, so the corresponding AMASS reference clip can be reloaded later for reward computation.

The construction process relies on the native PHC termination signal as the ground-truth failure criterion. No manually designed fall detector or handcrafted reset condition is introduced. During imitation rollout, randomized external forces are periodically applied to different body segments. These perturbations induce a broad spectrum of unstable configurations ranging from mild balance loss to imminent falls. If the PHC environment triggers a termination event, the preceding temporal window is retrospectively labeled as dangerous. This procedure is shown in Algorithm 1.

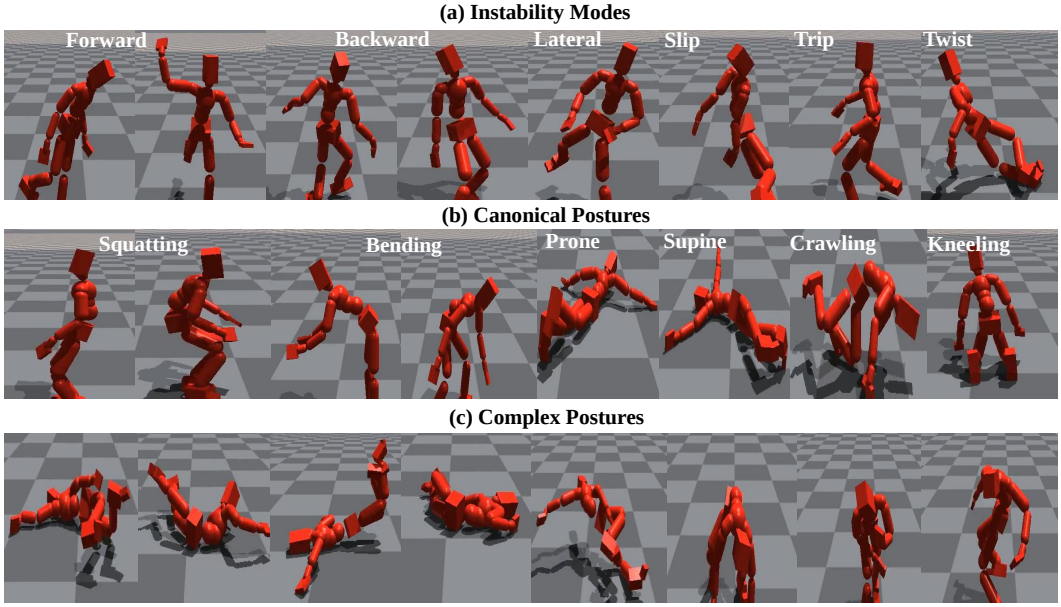


Figure 5: **Representative samples from the Realistic Motion-Driven Fall Dataset.** (a) Representative instability modes induced by randomized external perturbations, including forward, backward, lateral, slip, trip, and twist-induced loss of balance. (b) Canonical postures commonly used in the recovery process, including squatting, bending, prone, supine, crawling, and kneeling configurations. (c) Representative complex postures exhibiting highly articulated body configurations, unconventional support patterns, and severe instability.

Algorithm 1 Dangerous-State Dataset Construction

for each rollout episode **do**
  Sample an AMASS motion
  while episode not terminated **do**
    Execute action $a_t = \pi_{\text{PHC}}(o_t)$
    if push timer expires **then**
      Sample force magnitude f, duration d, target body b
      Apply external force to b
    end if
    if termination signal triggered **then**
      Label the previous T_{lookback} window as dangerous
    end if
    Label short post-push interval as dangerous
  end while
  Save rollout segment into dataset
end for

The coverage and dynamics of the collected dataset are shown in Figure 4. For the pose coverage analysis, features are embedded with t-SNE after standardization, using the state vectors (root state, DOF position, and DOF velocity) as input. As illustrated in Figure 4(a), multiple density centers are observed with notable interconnections rather than forming isolated clusters. This suggests that the dataset encompasses various fall modes (such as forward toppling, backward falling, lateral slipping, etc.), demonstrating a diverse and continuous coverage of postures. For the dynamics analysis, we computed the root angular velocity and plotted its empirical distribution, as shown in Figure 4(b). It is predominantly distributed between 0–5 rad/s, peaking around 0.5 rad/s.There is a sparse long tail extending up to 20, but the density is very low in the 5–20 range. This suggests that the collected states mainly cover moderate instabilities while still retaining rare extreme motions, which is desirable for robust recovery learning.

Table 1: Main results on recovery performance. SR: success rate; TTS: time to success; RCR: re-collapse rate; Upr.: uprightness. Lower is better for TTS, RCR, Energy, Torque Smooth, Norm. Jerk; higher is better for Upr.

Model	SR (%) ↑	TTS (s) ↓	RCR (%) ↓	Upr. ↑	Energy ↓	Torque Smooth ↓	Norm. Jerk ↓
PHC (Luo et al., 2023)	72.36	3.65	9.08	0.9855	0.2263	0.0625	1.39e+07
Heuristic + Tracking (Standing ref)	32.83	2.07	20.78	0.9885	0.0133	0.0081	9.72e+06
Heuristic + Tracking (AMASS ref)	90.54	2.79	4.38	0.9899	0.0356	0.0119	1.28e+07
Ours (Heuristic + Tracking + Regularization)	**98.34**	**2.77**	**3.12**	0.9886	0.0340	0.0116	1.24e+07

From the collected 2M+ dangerous frames, we subsample 26K+ frames based on uprightness, root height and linear/angular velocity, and then randomly split into training and test sets at an 8:2 ratio. This sampling ensures diverse recovery difficulties, with easy (upright > 0.7, 76.9%), medium ($0.3 <$ upright < 0.7, 8.7%), and hard (upright < 0.3, 14.5%).

Figure 5 presents samples from the proposed Realistic Motion-Driven Fall Dataset, which captures a diverse range of disturbance-induced dangerous states beyond the limited manually designed fallen poses in conventional recovery datasets. The dataset covers both structured and unstructured dangerous states, substantially enriching the diversity of recovery scenarios and training distributions.

4.2 Experimental Setup

Evaluation Metrics: We comprehensively evaluate the recovery policies across three dimensions: (1) Task performance, measured by success rate (SR), time-to-stand (TTS), and re-fall rate (RFR); (2) Postural stability, captured by termination uprightness to ensure balance; and (3) Hardware feasibility, evaluated via energy consumption, torque smoothness, and normalized jerk to penalize mechanical inefficiency and high-frequency control artifacts.

Implementation Details: The policy takes stacked proprioceptive frames initialized by Kinematic Prior History Injection (default history length $k = 5$) and outputs 69-D actions mapped to PD targets. The PPO actor-critic uses a six-hidden-layer SiLU MLP and fixed action standard deviation. All experiments run in Isaac Gym on a single RTX4090. The control policy is run at 30 Hz, and simulation runs at 60 Hz.The adaptive curriculum starts at difficulty 0.1 and updates every 100 epochs: difficulty increases when success rate exceeds 0.80 and decreases when it drops below 0.40 (step size 0.05, floor 0.10). Early termination is triggered after the agent maintains a successful upright state for 60 consecutive steps to improve sample efficiency.

4.3 Main Evaluations

As shown in Table 1, our model achieves the highest recovery success rate among all baselines. A purely heuristic-driven policy is prone to reward hacking and exhibits non-human behaviors (e.g., bouncing or oscillatory poses). A tracking-only baseline PHC (Luo et al., 2023) attains a success rate of 72.36%, while combining heuristic and tracking rewards improves it to 90.54%, indicating the Heuristic rewards effectively promoted the recovery process. Further incorporating kinematic regularization increases the success rate to 98.34%. We also compare two tracking references: a fixed standing pose and diverse AMASS motion clips. The fixed-reference variant achieves only 32.83% SR, whereas the AMASS-referenced model reaches 90.54%, indicating that a diverse motion library provides more informative AMP signals.

Beyond SR, our model also delivers better TTS and RCR than other methods (except the standing-reference variant, whose low SR makes other metrics less meaningful). Our model reduces TTS by 0.88s relative to PHC, suggesting more decisive and efficient policy outputs. It also achieves the lowest RCR at 3.12%, substantially below PHC (9.08%) and the AMASS tracking baseline (4.38%), demonstrating that the regularization reduces over-aggressive maneuvers and yields more stable recovery. In terms of hardware feasibility, our method records an energy expenditure of only 0.0340 and a torque smoothness of 0.0116, approximately $1/6$ of PHC. The normalized jerk is the lowest among all baselines.

Table 2: Ablation studies of Kinematic Prior History Injection (KPHI).

Setting	SR (%) ↑	TTS (s) ↓	RCR (%) ↓	Upr. ↑	Energy ↓	Torque Smooth ↓	Norm. Jerk ↓
Zero-padding	92.17	2.73	5.82	0.9901	0.0321	0.0114	1.27e+07
Frame duplication	92.56	**2.67**	5.36	0.9900	0.0347	0.0102	1.05e+07
Warmup(5 steps)	96.67	2.94	9.75	0.9898	0.0303	**0.0097**	**9.29e+06**
KPHI (**ours**)	**98.34**	2.77	**3.12**	0.9886	0.0340	0.0116	1.24e+07

Table 3: Ablations studies of heuristic annealing and regularization gating in phase-aware reward.

Setting	SR (%) ↑	TTS (s) ↓	RCR (%) ↓	Upr. ↑	Energy ↓	Torque Smooth ↓	Norm. Jerk ↓
w/o heuristic annealing	95.29	2.88	6.57	0.9882	0.0300	0.0100	1.12e+07
w/o regularization gating	92.80	2.78	5.31	0.9910	0.0308	0.0091	1.00e+07
Full model	**98.34**	**2.77**	**3.12**	0.9886	0.0340	0.0116	1.24e+07

4.4 Ablations and Analysis

Kinematic History initialization: We compare four history padding modes for dangerous frame resets, and results are shown in Table 2. Our Kinematic Prior History Injection (KPHI) achieves the highest success rate at 98.34%, outperforming zero-padding (92.17%), frame duplication (92.56%), and physics warmup (96.67%). Notably, the re-collapse rate (RCR) of our method is substantially lower than all other padding strategies. While our linear padding extrapolation is slightly inferior than Zero-padding and frame duplication in terms of TTS (by less than 0.1s) and does not achieve the lowest energy or smoothness compared with warmup, these differences are relatively small. By comparison, warmup strategy suffers from a substantially longer TTS of 2.94s and an RFT of 9.75%, while duplication and zero-padding yield success rates below 93%. Overall, our KPHI achieves the most favorable balance between robustness, recovery performance and motion quality.

Phase-Aware Reward: Heuristic annealing and regularization gating are ablated to verify the effectiveness of the phase-aware reward design. As shown in Table 3, both components contribute positively to recovery performance. Removing heuristic annealing decreases SR from $98.34\%$to 95.29%, indicating that the fixed recovery guide objective introduces optimization bias during late-stage recovery. The annealing strategy allows the policy to focus on stabilization and motion refinement. Similarly, removing regulation gating reduces SR to 92.80%, suggesting that applying stabilization-oriented regularization throughout the recovery process may hinder aggressive behaviors required during the early stages. The gating mechanism better balances recovery exploration and stabilization. Although removing either of these two parts results in a slight improvement in motion consistency metrics, it also restricts the exploration of effective recovery strategies, ultimately reducing recovery robustness. In general, the proposed full framework prioritizes successful recovery and long-term stability, achieving the highest success rate and the lowest re-collapse rate.

5 Conclusion

In this work, we focus on the recovery capabilities of humanoid robots in extremely dangerous situations. We present a unified framework integrating kinematic-prior history injection with phase-aware reward scheme, and train the policy using the diverse dangerous frame collected. Our method achieves a success rate of 98.34% with fast recovery (2.77 s TTS) and the lowest re-collapse rate (3.12%). The ablations further suggest that KPHI improves robustness to abrupt resets and that annealed heuristics plus gated regularization jointly balance aggressive exploration and stable standing.

Limitations: Although our **KiPAR** performed well in simulation, it is based on idealized proprioceptive sensing and contact models. Future work will focus on sim2real transfer, robustness under sensor noise and latency, while striving to achieve reliable recovery across various complex terrains.

AI USE STATEMENT

Generative AI tools were used during the preparation of this work to assist with writing and polishing, retrieval and discovery of relevant literature, coding and debugging. All AI-generated suggestions, code-related recommendations, and retrieved information were critically reviewed and verified by the authors. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work does not introduce new training data or collect human-subject data. The proposed method inherits potential risks and biases from the underlying pretrained generative models.

REPRODUCIBILITY STATEMENT

We provide the complete mathematical formulation, optimization procedure, and implementation details required to reproduce the proposed method. Hyperparameter settings, evaluation protocols, and additional experimental results are provided in the main paper and appendix. Code will be released upon publication.

REFERENCES

Trevor Louis Ablett. *Addressing Distribution Shift in Robotic Imitation Learning*. PhD thesis, University of Toronto (Canada), 2025.

Mazen Al Borno, Martin De Lasa, and Aaron Hertzmann. Trajectory optimization for full-body movements with complex contacts. *IEEE transactions on visualization and computer graphics*, 19(8):1405–1414, 2012.

Longbing Cao. Ai robots and humanoid ai: Review, perspectives and directions. *arXiv preprint arXiv:2405.15775*, 2024.

Penghui Chen, Yushi Wang, Changsheng Luo, Wenhan Cai, and Mingguo Zhao. Hifar: Multi-stage curriculum learning for high-dynamics humanoid fall recovery. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2908–2915. IEEE, 2025.

Kiyoshi Fujiwara, Fumio Kanehiro, Shuuji Kajita, Kazuhito Yokoi, Hajime Saito, Kensuke Harada, Kenji Kaneko, and Hirohisa Hirukawa. The first human-size humanoid that can fall over safely and stand-up again. In *Proceedings IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 2, pp. 1920–1926. IEEE, 2003.

Clément Gaspard, Marc Duclusaud, Grégoire Passault, Mélodie Daniel, and Olivier Ly. Frasa: An end-to-end reinforcement learning agent for fall recovery and stand up of humanoid robots. In *IEEE International Conference on Robotics and Automation*, pp. 15994–16000. IEEE, 2025.

Zhaoyuan Gu, Junheng Li, Wenlan Shen, Wenhao Yu, Zhaoming Xie, Stephen McCrory, Xianyi Cheng, Abdulaziz Shamsah, Robert Griffin, C Karen Liu, et al. Humanoid locomotion and manipulation: Current progress and challenges in control, planning, and learning. *IEEE/ASME Transactions on Mechatronics*, 31(2):2300–2330, 2026.

Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. *arXiv preprint arXiv:2406.08858*, 2024.

Tairan He, Wenli Xiao, Toru Lin, Zhengyi Luo, Zhenjia Xu, Zhenyu Jiang, Jan Kautz, Changliu Liu, Guanya Shi, Xiaolong Wang, et al. Hover: Versatile neural whole-body controller for humanoid robots. In *IEEE International Conference on Robotics and Automation*, pp. 9989–9996. IEEE, 2025a.

Xialin He, Runpei Dong, Zixuan Chen, and Saurabh Gupta. Learning getting-up policies for real-world humanoid robots. *arXiv preprint arXiv:2502.12152*, 2025b.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29, 2016.

Tao Huang, Junli Ren, Huayi Wang, Zirui Wang, Qingwei Ben, Muning Wen, Xiao Chen, Jianan Li, and Jiangmiao Pang. Learning humanoid standing-up control across diverse postures. *arXiv preprint arXiv:2502.08378*, 2025.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013.

Yuan Jiang, Zhiguang Cao, Yaoxin Wu, Wen Song, and Jie Zhang. Ensemble-based deep reinforcement learning for vehicle routing problems under distribution shift. *Advances in Neural Information Processing Systems*, 36:53112–53125, 2023.

Fumio Kanehiro, Kenji Kaneko, Kiyoshi Fujiwara, Kensuke Harada, Shuuji Kajita, Kazuhito Yokoi, Hirohisa Hirukawa, Kazuhiko Akachi, and Takakatsu Isozumi. The first humanoid robot that has the same size as a human and that can lie down and get up. In *IEEE International Conference on Robotics and Automation*, volume 2, pp. 1633–1639. IEEE, 2003.

Fumio Kanehiro, Kiyoshi Fujiwara, Hirohisa Hirukawa, Shin'ichiro Nakaoka, and Mitsuharu Morisawa. Getting up motion planning using mahalanobis distance. In *IEEE International Conference on Robotics and Automation*, pp. 2540–2545. IEEE, 2007.

Sotaro Katayama, Masaki Murooka, and Yuichi Tazaki. Model predictive control of legged and humanoid robots: models and algorithms. *Advanced Robotics*, 37(5):298–315, 2023.

Harin Kim, Donghyeon Seo, and Donghan Kim. Push recovery control for humanoid robot using reinforcement learning. In *IEEE International Conference on Robotic Computing*, pp. 488–492. IEEE, 2019.

Eric Krotkov, Douglas Hackett, Larry Jackel, Michael Perschbacher, James Pippine, Jesse Strauss, Gill Pratt, and Christopher Orlowski. The darpa robotics challenge finals: Results and perspectives. In *The DARPA robotics challenge finals: Humanoid robots to the rescue*, pp. 1–26. Springer, 2018.

Zhengyi Luo, Jinkun Cao, Kris Kitani, Weipeng Xu, et al. Perpetual humanoid control for real-time simulated avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10895–10904, 2023.

Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Sirui Chen, Fernando Castaneda, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, et al. Sonic: Supersizing motion tracking for natural humanoid whole-body control. *arXiv preprint arXiv:2511.07820*, 2025.

Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5442–5451, 2019.

Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021.

Gabriel B Margolis, Ge Yang, Kartik Paigwar, Tao Chen, and Pulkit Agrawal. Rapid locomotion via reinforcement learning. *The International Journal of Robotics Research*, 43(4):572–587, 2024.

Mayank Mittal, Pascal Roth, James Tigue, Antoine Richard, Octi Zhang, Peter Du, Antonio Serrano-Munoz, Xinjie Yao, René Zurbrügg, Nikita Rudin, et al. Isaac lab: A gpu-accelerated simulation framework for multi-modal robot learning. *arXiv preprint arXiv:2511.04831*, 2025.

Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions On Graphics*, 37(4):1–14, 2018.

Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics*, 40 (4):1–20, 2021.

Xue Bin Peng, Yunrong Guo, Lina Halper, Sergey Levine, and Sanja Fidler. Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters. *ACM Transactions On Graphics*, 41(4):1–17, 2022.

Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016.

Jerry Pratt, John Carff, Sergey Drakunov, and Ambarish Goswami. Capture point: A step toward humanoid push recovery. In *IEEE-RAS international conference on humanoid robots*, pp. 200–207. Ieee, 2006.

Giulio Romualdi, Stefano Dafarra, Giuseppe L'Erario, Ines Sorrentino, Silvio Traversaro, and Daniele Pucci. Online non-linear centroidal mpc for humanoid robot locomotion with step adjustment. In *International Conference on Robotics and Automation*, pp. 10412–10419. IEEE, 2022.

Max Schwenzer, Muzaffer Ay, Thomas Bergs, and Dirk Abel. Review on model predictive control: An engineering perspective. *International journal of advanced manufacturing technology*, 117 (5):1327–1349, 2021.

Qincheng Sheng, Zhongxiang Zhou, Jinhao Li, Xiangyu Mi, Pingyu Xiang, Zhenghan Chen, Haocheng Xu, Shenhan Jia, Xiyang Wu, Yuxiang Cui, et al. A comprehensive review of humanoid robots. *SmartBot*, 1(1):e12008, 2025.

Jörg Stückler, Johannes Schwenk, and Sven Behnke. Getting back on two feet: Reliable standing-up routines for a humanoid robot. In *IAS*, pp. 676–685, 2006.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *IEEE/RSJ international conference on intelligent robots and systems*, pp. 5026–5033. IEEE, 2012.

Duy Hoa Tran, Fred Hamker, and John Nassour. A humanoid robot learns to recover perturbation during swinging motion. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 50(10): 3701–3712, 2018.

Yinuo Wang, Yuanyang Qi, Jinzhao Zhou, Pengxiang Meng, and Xiaowen Tao. Humam: Humanoid motion control via end-to-end deep reinforcement learning with mamba. *arXiv preprint arXiv:2509.18046*, 2025.

Jinze Wu, Guiyang Xin, Chenkun Qi, and Yufei Xue. Learning robust and agile legged locomotion using adversarial motion priors. *IEEE Robotics and Automation Letters*, 8(8):4975–4982, 2023.

Zhengjie Xu, Ye Li, Kwan-yee Lin, and Stella X Yu. Unified humanoid fall-safety policy from a few demonstrations. *arXiv preprint arXiv:2511.07407*, 2025.

Chuanyu Yang, Can Pu, Guiyang Xin, Jie Zhang, and Zhibin Li. Learning complex motor skills for legged robot fall recovery. *IEEE Robotics and Automation Letters*, 8(7):4307–4314, 2023.

SUPPLEMENTARY MATERIAL

A REALISTIC MOTION-DRIVEN FALL DATASET CONSTRUCTION

A.1 EXTERNAL DISTURBANCE GENERATION

To maximize state diversity, external disturbances are randomly generated in terms of magnitude, duration, direction, and application location. The horizontal force direction is uniformly sampled from $[0, 2\pi)$, while the vertical component is independently sampled from a predefined range. Let f_h denote the horizontal magnitude and f_z the vertical component. The resulting force vector is

$$\mathbf{F} = \begin{bmatrix} f_h \cos\theta \\ f_h \sin\theta \\ f_z \end{bmatrix}, \qquad \theta \sim \mathcal{U}(0, 2\pi). \tag{S1}$$

To encourage diverse failure modes, perturbations are applied to different body regions with unequal probabilities: 60% to the pelvis/torso, 20% to the head, and 20% to the thigh/calf.Table S1 summarizes the default disturbance generation parameters.

Table S1: Default disturbance generation parameters.

Parameter	Symbol	Value
Horizontal force magnitude	f_h	$[350, 900]$ N
Vertical force component	f_z	$[-75, 75]$ N
Push interval	T_{int}	$[120, 260]$ steps
Push duration	T_{dur}	$[2, 6]$ steps
Lookback window	T_{lookback}	1.0 s
Immediate instability window	T_{unstable}	0.5 s

A.2 DANGEROUS-STATE LABELING

Termination-induced danger. When the environment emits a termination signal, all frames within the preceding lookback window

$$[t - T_{\text{lookback}},\, t] \tag{S2}$$

are assigned a positive danger label.

Post-push instability. Immediately after a disturbance is applied, a short instability interval is labeled as dangerous:

$$[t_{\text{push}},\, t_{\text{push}} + T_{\text{unstable}}]. \tag{S3}$$

This captures transient balance-loss behaviors that may not necessarily lead to a full fall.

Consequently, the dataset contains both imminent-failure states and recoverable unstable states, which is essential for learning robust recovery behaviors.

A.3 DATASET STRUCTURE

Each row in the dataset corresponds to a single simulator timestep. The stored information is shown in Table S2.

Table S2: Main fields stored in the Dangerous-State Dataset.

Category	Fields
Danger Labels	*Binary danger labels*
Observation	*root_height*, *local positions*, *local rotations*, *local velocities*, *local angular velocities*
Simulator State	*root_state_**, *dof_pos_**, *dof_vel_**
Reference Motion	*motion_id*, *motion_time*
Future Tracking Features	*reference body position offsets*, *reference velocity offsets*

B Details of Kinematic Prior History Injection

B.1 Observation Structure

The recovery policy adopts a stacked observation representation to capture temporal dynamics. Instead of feeding only the current observation, the policy receives a history window,

$$H_t = [\mathbf{o}_{t-k+1}, \mathbf{o}_{t-k+2}, \cdots, \mathbf{o}_{t-1}, \mathbf{o}_t] \tag{S4}$$

where k denotes the total number of frames, including the current frame, and each observation $\mathbf{o}_t = \Phi(\mathbf{x}_t)$ is generated by the environment observation encoder $\Phi(\cdot)$ from the simulator state $\mathbf{x}_t$.

During normal simulation, this history window is naturally accumulated as the robot evolves over time. However, immediately after an episode reset, only the current state$\mathbf{x}_t$ is available, while all previous observations are undefined. Therefore, the history buffer must be initialized before the first policy evaluation. Existing history padding strategies mainly include three approaches:

- **Zero-Padding** initializes every historical observation as an zero vector, which introduces a severe distribution mismatch because the policy has never observed such artificial histories during normal operation.
- **Frame duplication** directly duplicates the current observation for every history frame. Although computationally inexpensive, this initialization implicitly assumes that the robot has remained completely static throughout the history window. When the robot starts from a high-speed falling state, the stacked observations become inconsistent with the instantaneous velocity, resulting in an unrealistic temporal discontinuity.
- **Physics Warm-up** avoids synthetic history generation by executing several zero-action simulation steps before the first policy inference. The observation history is naturally accumulated from the simulator rollout. While this strategy preserves physical consistency, it delays policy intervention and may further deteriorate unstable recovery states before control begins.

KPHI synthesizes a short kinematic history from the current state before the first physics step. This approximation does not enforce contact or dynamic consistency.

B.2 History Buffer Construction

KPHI constructs the observation history through linear backward extrapolation. Given the sampled dangerous frame state is

$$\mathbf{x}_t = (\mathbf{p}_t, \mathbf{R}_t, \mathbf{v}_t, \boldsymbol{\omega}_t, \mathbf{q}_t, \dot{\mathbf{q}}_t) \tag{S5}$$

where $\mathbf{p}_t$ and $\mathbf{R}_t$ denote the root position and orientation, $\mathbf{v}_t$ and $\boldsymbol{\omega}_t$ are the root linear and angular velocities, $\mathbf{q}_t, \dot{\mathbf{q}}_t$ represent the joint positions and velocities.

we assume that the robot undergoes constant linear and joint velocities within the short history horizon. Under this first-order kinematic assumption, previous states can be reconstructed by linearly extrapolating the current state backward in time. For the j-th previous timestep,

$$\delta_j = j\Delta t, \qquad j = 1, \ldots, k-1, \tag{S6}$$

Table S3: Detailed phase-aware reward formulation.

Reward Term	Expression	Weight	Purpose
Root Height	$R_h = \text{clip}(h/0.9, 0, 1)$	$w_h(0.03)$	Encourage pelvis elevation
Upright Orientation	$R_u = (z_{\text{proj}} + 1)/2$	$w_u(0.12)$	Promote vertical torso alignment
Height Change (Upward Vel)	$R_v = 1 - \exp(-2\max(v_z, 0))$	$w_v(0.01)$	Encourage upward motion
Survival Reward	$R_s = \mathbb{I}[\text{not terminated}]$	$w_s(0.1)$	Prevent premature collapse
Action Magnitude Penalty	$R_{\text{action}} = w_{\text{action}} g_{\text{pen}} \|a_t\|^2$	$w_{\text{action}} = 0.005$	Penalize large actions
AMP Prior	$r_t^{\text{amp}} = -s_{\text{disc}} \log(\max\{1 - D(\xi_t), \epsilon\})$	$s_{\text{disc}} = 2,\ \epsilon = 10^{-4}$	Maintain natural motion prior
Root Point-Goal Tracking	$R_{pg} = k_{pg} \min(d_{t-1} - d_t,\ d_{max})$	$k_{pg}(9),\ d_{max}(1/3)$	Encourage root convergence to reference
Near-reference Imitation	$0.5\,\mathbb{I}[d_t \le 0.25\,\text{m}] R_{\text{imit}}$	0.5	Track synchronized reference
Power Penalty (task)	$R_{\text{power}} = -c_{\text{power}} \mathbb{I}[n_t > 3] \sum_i \lvert\tau_i \dot{q}_i\rvert$	$c_{\text{power}} = 0.00008$	Discourage high power
Action Rate Penalty	$R_{\text{rate}} = w_{\text{rate}} g_{\text{pen}} \|a_t - a_{t-1}\|^2$	$w_{\text{rate}} = 0.07$	Suppress abrupt control changes
Torque Penalty	$R_\tau = w_\tau g_{\text{pen}} \|\tau\|^2$	$w_\tau = 0.000006$	Penalize excessive actuation
Fall-back Penalty	$R_{\text{fall}} = w_{\text{fall}} b_{\text{drop},t}$	$w_{\text{fall}} = 20$	Penalize loss of standing
CoM-proxy Support Reward	$R_{\text{com}} = w_{\text{com}} g_{\text{com}} \max\{0, \tanh((r_{\text{sup}} + m - d_{\text{com}})/s)\}$	$w_{\text{com}} = 0.005$	Encourage stable support
Stable Standing Reward	$R_{\text{stable}} = w_{\text{stable}} g_{\text{stable}} e^{-k_{\text{lin}} \min\{\|v\|, c_v\}} e^{-k_{\text{ang}} \min\{\|\omega\|, c_\omega\}}$	$w_{\text{stable}} = 0.40$	Encourage momentum dissipation
Success Bonus	$R_{\text{success}} = R_{\text{terminal}} b_{\text{succ},t}$	$R_{\text{terminal}} = 10$	Reward confirmed recovery

root position is reconstructed as

$$\widehat{\mathbf{p}}_{t-j} = \mathbf{p}_t - \mathbf{v}_t \delta_j, \tag{S7}$$

while the joint configuration is approximated by

$$\widehat{\mathbf{q}}_{t-j} = \mathbf{q}_t - \dot{\mathbf{q}}_t \delta_j. \tag{S8}$$

For simplicity, the root orientation is assumed unchanged, and all velocity quantities are copied

$$\widehat{\mathbf{R}}_{t-j} = \mathbf{R}_t, \quad \widehat{\mathbf{v}}_{t-j} = \mathbf{v}_t, \quad \widehat{\boldsymbol{\omega}}_{t-j} = \boldsymbol{\omega}_t, \quad \widehat{\dot{\mathbf{q}}}_{t-j} = \dot{\mathbf{q}}_t. \tag{S9}$$

Each reconstructed state is temporarily written into the simulator, and the observation encoder is invoked to recompute the corresponding observation.

$$\widehat{\mathbf{o}}_{t-j} = \Phi(\widehat{\mathbf{x}}_{t-j}), \qquad H_t = (\widehat{\mathbf{o}}_{t-k+1}, \ldots, \widehat{\mathbf{o}}_{t-1}, \mathbf{o}_t). \tag{S10}$$

These observations are sequentially inserted into the history buffer from the oldest frame to the most recent one. After all historical observations have been generated, the simulator state is restored to the original reset state before physics simulation proceeds.

B.3 Reward Details

The recovery controller uses $R_{\text{total}} = 0.5R_{\text{env}} + 0.5r_t^{\text{amp}}$. Reference-conditioned tracking is part of R_{env}; the discriminator reward is computed and mixed separately. Only the height, uprightness and upward-velocity terms are multiplied by $\beta(\alpha)$; survival is not annealed.

Table S3 lists the active terms and coefficients. Expressions include their weights unless explicitly shown as unweighted component features (R_h, R_u, R_v, R_s). Positive penalty magnitudes are subtracted through R_{reg}; R_{power} is already negative and is added to R_{track}.

The gate parameters $(h_j^-, h_j^+, z_j^-, z_j^+)$ are $(0.55, 0.85, 0.78, 0.90)$ for g_{pen}, $(0.72, 0.88, 0.82, 0.92)$ for g_{com}, and $(0.75, 0.90, 0.75, 0.90)$ for g_{stable}; heights are in meters. The support reward uses $m = 0.04\,\text{m}$ and $s = 0.10\,\text{m}$. The stable-standing reward uses $k_{\text{lin}} = 1$, $k_{\text{ang}} = 0.7$, $c_v = 2.5\,\text{m/s}$ and $c_\omega = 2.5\,\text{rad/s}$. These values describe the inspected full-model configuration.

Let $b_{\text{drop},t}$ indicate the implemented fall-back event: the robot has previously satisfied $h > 0.82\,\text{m}$ and $z_{\text{proj}} > 0.85$ for 25 consecutive steps, subsequently violates that standing condition, and has no active penalty cooldown. A penalty disarms this event until standing is established again and starts a 20-step cooldown. Let $b_{\text{succ},t}$ indicate the success event used for early termination. With the time-bonus coefficient set to zero, $R_{\text{success}} = 10 b_{\text{succ},t}$. The power penalty is disabled when the environment progress counter $n_t \le 3$.

Table S4: PPO training hyperparameters.

Parameter	Value
Hidden layers	[2048, 1536, 1024, 1024, 512, 512]
Activation	SiLU
Discount factor γ	0.99
GAE λ	0.95
Learning rate	2×10^{-5}
Learning rate schedule	Constant
Rollout horizon	32
Mini-epochs	6
Minibatch size	16384
PPO clip ϵ	0.2
Critic loss coefficient	5.0
Entropy coefficient	0.0
Gradient clipping	50.0
Input normalization	Enabled
Value normalization	Enabled

C Implementation Details

C.1 Details of Recovery Policy

The recovery policy adopts the `amp_pnn` network implementation provided by PHC. During training, the progressive neural network (PNN) module is disabled (`has_pnn=False`), resulting in a standard feed-forward multilayer perceptron. The actor consists of six fully connected layers with hidden dimensions $[2048,\ 1536,\ 1024,\ 1024,\ 512,\ 512]$, using the SiLU activation function throughout. Continuous actions are modeled by a Gaussian policy with a fixed log standard deviation initialized to -2.9.

Table S4 summarizes the PPO optimization hyperparameters used throughout training.

C.2 Curriculum-Based Policy Training

Difficulty-aware sampling: Each dangerous snapshot is assigned a difficulty score based on uprightness. From the root orientation quaternion (x, y, z, w) we compute the vertical projection $z_{\text{proj}} = 1 - 2(x^2 + y^2)$, and map it to difficulty

$$d = 0.5\,(1 - z_{\text{proj}}), \tag{S11}$$

where $d \approx 0$ corresponds to near-upright states and $d \approx 1$ to upside-down states. During reset, we filter fall states by $d \leq \tau$, where τ is the curriculum difficulty.

C.3 Adaptive Curriculum Update

To progressively increase the recovery difficulty, we maintain the cumulative numbers of successful recoveries and episode resets, and update the curriculum parameter τ every 100 training epochs. Let

$$r = \frac{N_{\text{success}}}{N_{\text{reset}}}, \tag{S12}$$

denote the current recovery success rate. The curriculum parameter is then updated according to

$$\tau \leftarrow \begin{cases} \min(\tau + 0.05,\ 1.0), & r > 0.80, \\ \max(\tau - 0.02,\ 0.10), & r < 0.40, \\ \tau, & \text{otherwise.} \end{cases} \tag{S13}$$

As the policy performance improves, increasingly challenging initial recovery states are gradually introduced, while poor performance automatically reduces the difficulty. This adaptive curriculum prevents abrupt transitions to highly difficult states and mitigates catastrophic forgetting during training.

Early-stop success criterion: We decouple success from tracking error and use purely physical criteria. A trajectory is considered successful if the root height exceeds $0.8\,\mathrm{m}$ and uprightness satisfies $z_{\text{proj}} > 0.82$ for H consecutive steps (default $H = 60$). Once a success streak is reached, we terminate the episode early during training.

C.4 Domain Randomization

Domain randomization is applied only during training to improve robustness against modeling inaccuracies and facilitate Sim2Real transfer. The randomized quantities include physical properties, actuator parameters, control latency, and observation noise. Each parameter is independently sampled from a uniform distribution over its predefined interval at the beginning of every training episode.

Specifically, we randomize actor collision-shape friction offsets, link mass multipliers, PD joint gain multipliers, and control delay steps at reset. Additive channel-specific observation noise is sampled at each observation. Domain randomization is confined to training; evaluation uses nominal physical parameters, zero additional delay and no observation noise. Table S5 summarizes the randomized parameters together with their corresponding sampling ranges used throughout training.

Table S5: Domain Randomization Parameters

Parameter	**Range / Scale**
Ground Friction Offset	$[-0.6, 1.2]$
Link Mass Multiplier	$[0.7, 1.3]$
K_p Gain Multiplier	$[0.75, 1.25]$
K_d Gain Multiplier	$[0.75, 1.25]$
Control Delay Steps	$[1, 3]$ steps
Global Observation Noise Level	1.0
Linear Velocity Noise Scale	0.2
Angular Velocity Noise Scale	0.5
Gravity Vector Noise Scale	0.1
Body Position Noise Scale	0.01