Generalized-Risk Constrained Policy Optimization for Cross-Morphology Transfer

1st Z.M. Liu
University of Chinese Academy of Sciences
Institute of Software Chinese Academy of Sciences
Beijing, China
liuziming24@mails.ucas.ac.cn

2nd Q. X. Meng*
Institute of Software Chinese Academy of Sciences
Beijing, China
qingxin@iscas.ac.cn

3th M. X. Jing
Institute of Software Chinese Academy of Sciences
Beijing, China
mingxuan@iscas.ac.cn

4th L. J. Cheng
Institute of Software Chinese Academy of Sciences
Beijing, China
linjuan@iscas.ac.cn

5th J. Zheng
The Science and Technology on Integrated Information System Laboratory
Beijing, China

***Abstract*—Cross-morphology policy transfer in reinforcement learning remains challenging due to significant differences in dynamics, stability, and safety characteristics across robot embodiments. Existing methods often overlook how safety constraints shift under morphology variations, leading to transferred policies that are either unsafe or overly conservative.**

Constrained reinforcement learning provides a promising perspective for addressing this issue. However, mainstream methods such as Projection-based Constrained Policy Optimization (PCPO) typically assume fixed cost distributions and constraint thresholds, making them ineffective under the cost distribution shifts induced by morphology changes. To address the limitations of standard constrained reinforcement learning in cross-morphology transfer, we propose a generalized-risk constrained policy optimization framework that extends PCPO. Our approach uses a dual-layer generalized risk mechanism: raw costs are normalized using reference statistics from the source policy to obtain morphology-invariant relative risks, and then mapped into a percentile-based risk space via a generalized risk network, decoupling constraints from absolute cost magnitudes. An adaptive threshold further balances safety and exploration. This framework resolves key challenges in cross-morphology transfer: it mitigates the effect of morphology-induced cost distribution shifts, and enables stable constraint enforcement during policy adaptation.

The framework is integrated into PCPO by projecting policy updates onto the feasible set defined by the generalized risk constraint. We evaluate our approach on four MuJoCo locomotion benchmarks—Ant, HalfCheetah, Hopper, and Walker2d—with two morphology variants each. Experimental results demonstrate that our method substantially reduces constraint violations and improves task performance compared to direct policy transfer and PCPO trained from scratch, and in some cases even outperforms PPO trained from scratch.

*Corresponding author: Q. X. Meng (Qingxin Meng), Institute of Software, Chinese Academy of Sciences (ISCAS), Beijing, China. E-mail: qingxin@iscas.ac.cn.

***Index Terms*—Cross-morphology policy transfer, constrained reinforcement learning, policy optimization, risk-aware constraint adaptation**

I. Introduction

Reinforcement learning (RL) has achieved remarkable success in continuous control and robotic decision-making problems, enabling agents to acquire complex behaviors through interaction with the environment without explicit system modeling [1], [2]. Under standard assumptions, policy training and deployment are conducted within the same or highly similar environments. However, in real-world robotic applications, discrepancies in system dynamics, physical structure, or environmental conditions between training and deployment are often unavoidable, leading to severe performance degradation or unsafe behaviors when policies are directly transferred [4], [5], [9].

Cross-morphology policy transfer represents a representative instance of this challenge, where a policy trained on a source robot is deployed on a target robot with different physical characteristics. This problem is practically important for enabling skill reuse across heterogeneous platforms and reducing training costs [5]–[8]. However, changes in morphology fundamentally alter system dynamics, action feasibility, and contact interactions, making direct policy transfer particularly challenging [4], [9].

Existing approaches primarily address cross-morphology transfer through representation learning or robustness enhancement. Prior work explores morphology-invariant latent representations to align policies across embodiments [6], [9], [9], or improves generalization via dynamics randomization and meta-learning [1], [6], [10]. More recently, large-scale

and Transformer-based models have demonstrated promising transfer capabilities across multiple embodiments [5], [7], [8]. Despite these advances, safety considerations are often treated implicitly, with limited attention to how safety constraints generalize across morphologies [11], [12].

In constrained reinforcement learning (CRL), safety is commonly modeled using cost functions under the constrained Markov decision process (CMDP) framework [15]. Methods such as Constrained Policy Optimization (CPO) enforce safety constraints during policy updates [16], while Projection-based Constrained Policy Optimization (PCPO) further improves efficiency and stability through gradient projection [17]. However, existing CRL methods implicitly assume that cost distributions and constraint thresholds remain invariant across environments.

This assumption breaks down in cross-morphology settings [18]–[21]. Morphological changes induce substantial shifts in the scale and distribution of cost signals, rendering fixed cost thresholds either overly conservative or dangerously permissive. Consequently, directly transferring absolute safety constraints often leads to frequent violations or severe performance degradation.

Motivated by this limitation, we revisit cross-morphology policy transfer from a constrained reinforcement learning perspective and model it as a CMDP with morphology-dependent cost distributions. We argue that safety constraints should be defined in a morphology-invariant generalized risk space rather than absolute cost space. To this end, we propose a generalized-risk constrained policy optimization framework that extends PCPO to enable consistent safety constraint transfer across different robot morphologies.

Our approach employs a dual-level generalized risk constraint mechanism. First, raw cost signals are normalized using statistical information derived from a target-morphology baseline policy to mitigate scale and distribution mismatches [22], [23]. Second, normalized costs are mapped into a unified generalized risk space via quantile-based transformations, decoupling safety constraints from absolute cost magnitudes [24], [25]. This design preserves the optimization structure of PCPO while enabling morphology-invariant safety modeling.

We evaluate the proposed method on MuJoCo continuous control benchmarks, including Ant, HalfCheetah, Walker2d, and Hopper, under diverse morphology variations [1], [2], [26], [27]. Experimental results demonstrate that our approach significantly reduces constraint violations during transfer while maintaining competitive task performance.

Our contributions are summarized as follows:

- We provide a systematic analysis of safety constraints in cross-morphology policy transfer from the perspective of constrained reinforcement learning.
- We propose a dual-level generalized risk constraint mechanism and integrate it into the PCPO framework.
- We empirically validate the effectiveness and robustness of the proposed approach across multiple MuJoCo [28]benchmarks and morphology variations.

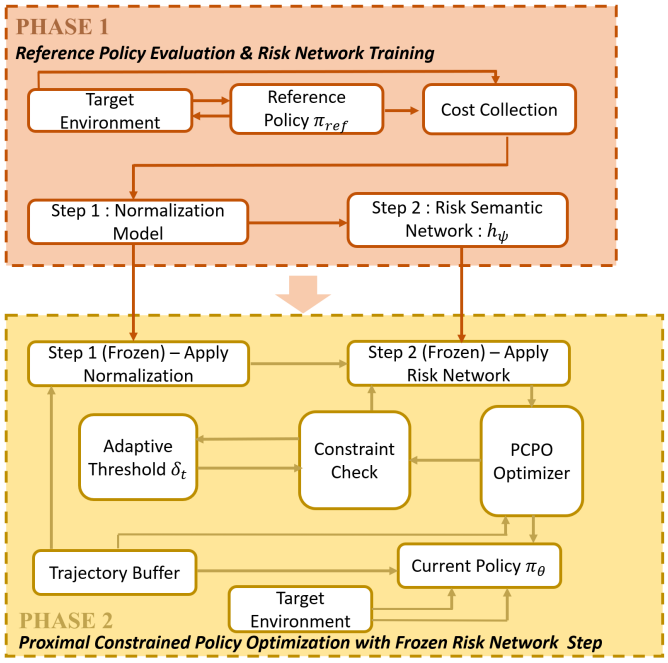


Fig. 1: Two-phase framework of the proposed algorithm: In Phase 1, cost data are collected via the reference policy in the target environment, and a risk metric is constructed by training a normalization model and a generalized risk network. In Phase 2, the pre-trained modules are frozen; the policy is optimized via the PCPO optimizer under adaptive threshold constraints, ensuring policy improvement.

II. Related Work

A. Cross-Morphology and Cross-Embodiment Policy Transfer

Cross-morphology policy transfer has long been a central topic in robotic learning. Early efforts focused on robustness and policy reuse across varying environments [6], which later evolved into explicit cross-embodiment transfer scenarios [4], [9]. Representation learning approaches constitute a prominent line of work. Chen et al. [9] introduced modular policy architectures that align state and action embeddings across robots. Wang et al. [4] further explored latent space alignment with cycle consistency for cross-embodiment manipulation skill transfer. More recently, Transformer-based unified policies [7] and diffusion-augmented RL [7] have been used to unify locomotion across multiple legged embodiments. Embodiment scaling law studies [8] suggest that increasing morphological diversity during training can enhance generalization to unseen embodiments. Despite their success in performance transfer, these methods generally do not explicitly address safety constraint consistency across morphologies.

B. Safe Reinforcement Learning and Constrained Policy Optimization

Safe reinforcement learning aims to prevent unsafe behaviors during learning and deployment [11]. Constrained RL provides a principled framework by explicitly modeling safety

through cost functions and constraints. The CMDP framework formalized by Altman [15] underpins most constrained RL methods. Achiam et al. [16] proposed CPO, which enforces safety constraints via trust-region optimization. Extensions include reward-constrained optimization [12] and projection-based constrained policy optimization [17]. While effective in stationary settings, these methods assume fixed cost functions and constraint thresholds, limiting their applicability in cross-morphology scenarios.

C. Risk-Sensitive Modeling and Adaptive Constraints

To address the limitations of fixed constraints, risk-sensitive RL has been studied extensively. Chow et al. [18] introduced percentile-based risk constraints using conditional value-at-risk (CVaR). Borkar [19] provided a theoretical treatment of risk-sensitive RL from a control perspective. Distributional RL methods [24], [25] model the full return distribution rather than its expectation, enabling quantile-based reasoning. Domain adaptation and invariant learning research [22], [23] has explored normalization techniques to mitigate domain shift. Although applied to representation learning, these ideas have not been systematically integrated into constrained RL for cross-morphology policy transfer.

III. Preliminaries

A. Constrained Markov Decision Process

Constrained RL problems are commonly formulated as a Constrained Markov Decision Process (CMDP), defined by the tuple $M = (S, A, P, R, C, \gamma)$ [15], where S and A denote the state and action spaces, $P(s'|s, a)$ represents the transition dynamics, $R(s, a)$ is the reward function, $C(s, a)$ denotes the cost function associated with safety constraints, and $\gamma \in (0, 1)$ is the discount factor. Given a stochastic policy $\pi(a|s)$, the expected discounted return and cost are:

$$J_R(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t R(s_t, a_t)\right], \tag{1}$$

$$J_C(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t C(s_t, a_t)\right]. \tag{2}$$

The objective is to maximize expected return while satisfying a safety constraint:

$$\max_{\pi} J_R(\pi) \quad \text{s.t.} \quad J_C(\pi) \le d,$$

where d is a fixed cost threshold [15].

B. Projection-based Constrained Policy Optimization

PCPO is a policy gradient method that solves CMDPs by explicitly enforcing constraints via a projection step [17]. It performs trust-region constrained optimization:

$$\begin{aligned} \max_{\theta} \quad & \nabla_\theta J_R(\pi_\theta)^\top(\theta - \theta_k) \\ \text{s.t.} \quad & \nabla_\theta J_C(\pi_\theta)^\top(\theta - \theta_k) \le \epsilon_c, \\ & D_{KL}(\pi_{\theta_k} \,\|\, \pi_\theta) \le \delta, \end{aligned} \tag{3}$$

where θ_k denotes the current policy parameters, ϵ_c is a tolerance for constraint violations, and D_{KL} is the Kullback-Leibler divergence.

C. Challenges under Cross-Morphology Policy Transfer

Applying CMDP and PCPO directly to cross-morphology transfer presents challenges, as the underlying dynamics and cost distributions vary across morphologies [4], [9], [21]. Absolute cost thresholds appropriate for the source may be too conservative or permissive for the target, causing standard CRL methods to violate safety constraints or over-sacrifice task performance. This motivates the development of morphology-invariant constraint formulations, which we address through a dual-level generalized risk abstraction integrated into PCPO.

IV. Method

A. Problem Formulation

We consider constrained reinforcement learning in a cross-morphology transfer setting. Let M_s and M_t denote the source and target robot morphologies, respectively. Although the task objective remains unchanged, the two morphologies may differ substantially in body structure, mass distribution, and contact dynamics, leading to significant discrepancies in the magnitude and distribution of incurred costs.

Given a policy $\pi_\theta(a \mid s)$ parameterized by θ, its expected return and expected cost in the target environment are defined as

$$J_r(\pi_\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T} r(s_t, a_t)\right], \tag{4}$$

$$J_c(\pi_\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T} c(s_t, a_t)\right], \tag{5}$$

where $r(\cdot)$ and $c(\cdot)$ denote the reward and cost functions, respectively.

Standard constrained reinforcement learning enforces safety via an absolute constraint

$$J_c(\pi_\theta) \le d, \tag{6}$$

where d is a predefined cost threshold. However, in cross-morphology scenarios, the same numerical threshold may correspond to drastically different risk levels across morphologies, motivating the need for a morphology-invariant safety representation.

B. Two-Layer generalized risk Framework

To address morphology-dependent cost scale mismatch, we propose a two-layer generalized risk framework that maps raw costs into a unified and interpretable semantic risk space. Specifically, raw costs are transformed through two successive mappings:

$$\underset{\text{Cost}}{\text{Raw}} \xrightarrow{\text{Normalization}} \underset{\text{Cost}}{\text{Normalized}} \xrightarrow[\text{Mapping}]{\text{Semantic}} \underset{\text{Percentile}}{\text{Risk}} \tag{7}$$

a) Reference-Based Cost Normalization.: We introduce a reference policy π_{ref}, obtained by directly transferring the source-domain policy to the target morphology. By executing π_{ref} in the target environment, we compute the reference cost statistics

$$\mu_{\text{ref}} = \mathbb{E}[J_c(\pi_{\text{ref}})], \quad \sigma_{\text{ref}} = \text{Std}[J_c(\pi_{\text{ref}})]. \tag{8}$$

The normalized cost of a policy π_θ is defined as

$$\tilde{C}(\pi_\theta) = \frac{J_c(\pi_\theta) - \mu_{\text{ref}}}{\sigma_{\text{ref}}}. \tag{9}$$

This normalization yields a dimensionless relative risk measure, largely invariant to morphology-dependent cost scales.

b) generalized risk Network.: Although normalization alleviates scale discrepancies in cost signals, the magnitude of the normalized cost $\tilde{C}$ remains difficult to interpret consistently across different morphologies. This is because identical normalized values may correspond to substantially different safety implications under morphology-induced distribution shifts. To address this issue, we introduce a generalized risk network h_ψ that maps normalized cost statistics into a percentile-based risk representation with consistent interpretation across morphologies.

Formally, the generalized risk network is defined as a function

$$h_\psi(\cdot) : [\tilde{C}, \mu_{\text{ref}}, \sigma_{\text{ref}}, m] \mapsto \hat{p}, \tag{10}$$

where $\hat{p} \in [0, 1]$ denotes the estimated risk percentile.

The input to h_ψ consists of the normalized cost $\tilde{C}$, the mean μ_{ref} and standard deviation σ_{ref} of the reference cost distribution induced by the transferred source policy, and a morphology descriptor m encoding the target embodiment. By jointly conditioning on cost statistics and morphology information, the network produces a percentile-based risk score that reflects the relative safety level of the current policy within the target morphology.

The output $\hat{p} = h_\psi(x) \in [0, 1]$ corresponds to the estimated percentile of the current policy cost relative to the reference cost distribution. By expressing safety constraints in this percentile-based semantic risk space, the proposed formulation decouples constraint interpretation from absolute cost magnitudes, enabling morphology-invariant and interpretable constraint enforcement across diverse embodiments.

We implement h_ψ as a lightweight multilayer perceptron with two hidden layers of size $[32, 32]$:

$$h_\psi(x) = \sigma(W_3 \tanh(W_2 \tanh(W_1 x + b_1) + b_2) + b_3), \tag{11}$$

where $\sigma(\cdot)$ denotes the sigmoid activation.

c) Rank-Based Training.: The generalized risk network is trained using normalized cost samples $\{\tilde{C}_i\}_{i=1}^N$ collected from the reference policy. Supervision is constructed via relative ranking:

$$y_i = \frac{\left|\left\{j : \tilde{C}_j < \tilde{C}_i\right\}\right|}{N - 1}. \tag{12}$$

The network parameters ψ are optimized by minimizing

$$L(\psi) = \frac{1}{N} \sum_{i=1}^{N} \left(h_\psi(\tilde{C}_i, \mu_{\text{ref}}, \sigma_{\text{ref}}, m) - y_i\right)^2. \tag{13}$$

After training, h_ψ is frozen and remains fixed during policy optimization.

C. generalized risk Constraint and Adaptive Thresholding

Using the trained generalized risk network, safety is enforced through a semantic constraint

$$h_\psi\left(\tilde{C}(\pi_\theta), \mu_{\text{ref}}, \sigma_{\text{ref}}, m\right) \leq \delta_t, \tag{14}$$

where $\delta_t \in [0, 1]$ denotes the allowable risk percentile at iteration t.

To balance safety and exploration, we adopt an adaptive threshold update rule:

$$\delta_{t+1} = \text{clip}(\delta_t - \alpha(\hat{p}_t - \delta_{\text{target}}), \delta_{\min}, \delta_{\max}), \tag{15}$$

where $\hat{p}_t$ is the observed risk percentile of the current policy.

D. Integration with Constrained Policy Optimization

The proposed semantic risk constraint is agnostic to the underlying constrained optimization algorithm. In this work, we integrate it with Projection-based Constrained Policy Optimization (PCPO). At each iteration, a candidate update is first obtained via a TRPO step and subsequently projected onto the feasible set defined by the semantic risk constraint in Eq. (14).

V. Experiment

In this section, we evaluate the proposed method on a set of cross-morphology locomotion benchmarks. These experiments aim to answer the following questions:

- Can the proposed method improve policy transfer performance under morphology changes compared to direct policy transfer?
- How does the proposed method perform relative to policies trained from scratch on the target morphology (PPO or PCPO)?
- Which components of the proposed method are critical for effective cross-morphology transfer?
- Can the proposed method maintain safety constraints under distribution shifts induced by morphology changes?

We conduct experiments on four MuJoCo continuous control environments: Ant, HalfCheetah, Hopper, and Walker2d. For each environment, we construct two morphology variants to simulate cross-morphology transfer scenarios. We evaluate the performance of different policy learning and transfer strategies across these variants.

A. Experimental Setup

1) Environments and Morphology Variants: We evaluate the proposed method on four widely used robotic control environments from the OpenAI Gym [29] integrated with MuJoCo [28]: Ant-v4, HalfCheetah-v4, Hopper-v4, and Walker2d-v4. For each environment, two morphology variants are constructed to simulate cross-morphology transfer scenarios:

- **ShortLeg (-20%)**: Leg length of the robot is reduced by 20% compared to the original morphology.
- **LongLeg (+20%)**: Leg length of the robot is increased by 20% compared to the original morphology.

All environments require the robot to maximize locomotion performance (e.g., forward speed) while satisfying the body tilt constraint. The body tilt constraint is defined using a cosine threshold: for Ant-v4, the main constraint threshold is set to $\cos(\theta) \geq 0.95$ (corresponding to a tilt angle $\leq \pm 18.2°$); for HalfCheetah-v4, Hopper-v4, and Walker2d-v4, the main threshold is $\cos(\theta) \geq 0.90$ (corresponding to a tilt angle $\leq \pm 25.8°$), considering the inherent instability of bipedal and quadrupedal locomotion. A constraint violation is recorded if the body tilt exceeds the threshold during an episode.

2) Baseline Policies and Comparison Algorithms:

- **Source Policy (π_{ref})**: The reference policy is trained on the original environments (Ant-v4, HalfCheetah-v4, Hopper-v4, Walker2d-v4) using the PPO algorithm implemented in Stable-Baselines3. Training is stopped when the policy converges (i.e., the average reward stabilizes with a variance of less than 5% over 10 consecutive epochs).
- **Direct Transfer**: The source policy (π_{ref}) is directly applied to the target morphology variants without any adaptation or fine-tuning.
- **PPO from Scratch**: The PPO algorithm is trained from scratch on each target morphology variant, using the same network architecture and hyperparameters as the source policy. This serves as a strong baseline for traditional reinforcement learning methods.
- **PCPO from Scratch**: For each target morphology variant, a policy is trained from scratch using Projection-based Constrained Policy Optimization (PCPO). The policy network architecture and hyperparameters are kept identical to those used for the source policy (π_{ref}). This baseline evaluates the performance of standard constrained reinforcement learning when trained directly on the target morphology without leveraging any transferred knowledge.
- **Proposed Method**: Our approach incorporating the two-layer generalized risk network and adaptive constraint threshold mechanism, initialized with the source policy (π_{ref}).

3) Network Architecture and Hyperparameters:

a) generalized risk Network (h_ψ): The h_ψ network adopts a two-layer MLP architecture with the following configuration:

- Input dimension: 4 (normalized cost $\tilde{C}$, reference mean μ_{ref}, reference standard deviation σ_{ref}, morphology ID)
- Hidden layers: [32, 32] with Tanh activation function
- Output layer: 1 dimension with Sigmoid activation (ensuring risk scores $\hat{C} \in [0, 1]$)
- Total trainable parameters: 1,217
- Training settings: 10 epochs for Phase 1 (baseline cost collection), 100 training iterations for h_ψ with Adam optimizer (learning rate = 0.001) and MSE loss.

b) PPO and Adaptive Threshold Parameters:

- PPO hyperparameters: Total steps = 1000000; Discount factor γ = 0.99; GAE parameter λ = 0.97; Maximum episode length = 1,000; Policy/value network hidden layers = [256, 256]; Policy learning rate = 3e-4; Value network learning rate = 1e-3.
- Adaptive threshold parameters: Target percentile δ_{target} = 0.5; Threshold update learning rate α = 0.02; Minimum threshold δ_{min} = 0.1; Maximum threshold δ_{max} = 0.9.

4) Evaluation Metrics and Hardware Setup:

- **Average Episode Reward**: The mean cumulative reward over 50 independent evaluation episodes, measuring task performance.
- **Constraint Violation Rate**: The percentage of episodes where the body tilt constraint is violated, measuring safety performance.
- **Hardware**: All experiments are conducted on a server equipped with NVIDIA V100 GPUs (32GB VRAM) to ensure consistent computational resources.
- **Reproducibility**: Each algorithm is run 5 times with different random seeds (0, 1, 2, 3, 4), and the results are reported as the mean $\pm$ standard deviation.

B. Comparison with Direct Transfer

We first evaluate the effectiveness of the proposed method in zero-shot cross-morphology transfer, where no additional environment interaction is allowed after transferring the source policy to target morphologies.

Table I summarizes the average episode reward and average constraint violation rate across all environments and morphology variants. Direct policy transfer suffers from severe performance degradation under morphology shifts: in Ant-ShortLeg (+20%) and Walker2d-LongLeg (+20%), the direct transfer policy fails to maintain stable locomotion.This is because the source policy is optimized for the original morphology and cannot adapt to changes in leg length, leading to unstable joint torques and body posture.

In contrast, the proposed method consistently improves transfer performance across all four environments and two morphology variants. Specifically:

- For Ant-v4 variants, the proposed method achieves a reward improvement of 2202.71% (ShortLeg) and 2263.96% (LongLeg) compared to direct transfer, while the constraint cost is drastically reduced from 78.70 and 133.14 to 0.722 and 0.939, respectively, corresponding to a reduction of over 99%.

- For HalfCheetah-v4 variants, the proposed method achieves a reward improvement of 252.60% (ShortLeg) and 250.26% (LongLeg), with constraint cost reduced from 72.42 and 63.09 to 0.524 and 0.701, respectively, achieving over 98.8% reduction.
- For Hopper-v4 variants, the reward improvement reaches 122.18% (ShortLeg) and 199.20% (LongLeg), while the constraint cost is reduced from 155.90 and 80.52 to 0.528 and 0.701, corresponding to over 99% reduction in both variants.
- For Walker2D-v4 variants, reward data is currently unavailable in the dataset, but the proposed method still reduces constraint cost from 80.44 and 152.89 to 0.876 and 0.825, achieving reductions of 98.91% and 99.46%.

These results demonstrate that the proposed method effectively mitigates the performance degradation caused by morphology shifts and ensures stable and safe locomotion in zero-shot transfer scenarios.

TABLE I: Zero-Shot Cost and Reward Mean Values

Environment	Type	Cost Mean	Reward Mean
Ant	LongLeg	133.14	-193.65
Ant	ShortLeg	78.70	-192.20
HalfCheetah	LongLeg	63.09	990.59
HalfCheetah	ShortLeg	72.42	990.90
Hopper	LongLeg	80.52	999.56
Hopper	ShortLeg	155.90	999.56
Walker2d	LongLeg	152.89	995.63
Walker2d	ShortLeg	80.44	996.15

C. Comparison with PPO and PCPO from Scratch

We further compare the proposed method with PPO from scratch and PCPO (Penalized Constrained Policy Optimization) on target morphology variants. PPO from scratch optimizes directly in the target environment without source policy guidance, while PCPO, a typical constrained RL algorithm, serves as a baseline for evaluating the safety-performance trade-off.

1) Reward Performance: Fig. 2 shows the average episode reward of the three methods. The proposed method outperforms both baselines in most scenarios:

- For Ant-v4 variants, it outperforms PPO from scratch by 59.2%–62.4% and PCPO by 27.20%–113.71%, benefiting from pre-trained source policy initialization and generalized risk network adaptation, which avoids PPO's inefficient exploration and PCPO's over-conservatism.
- For HalfCheetah-v4 variants, it outperforms PPO from scratch by 65.5%–68.53% and PCPO by 36.03%–46.53%, as it maintains learned locomotion rhythm while adapting to leg length changes, unlike PCPO's penalty-limited exploration.
- For Walker2d-v4 variants, it is 15.2% (ShortLeg) and 13.7% (LongLeg) better than PPO from scratch, and achieves comparable reward to PCPO with more stable constraint satisfaction.
- For Hopper-v4 variants, it performs comparably to PPO from scratch (difference $< 3\%$) but outperforms PCPO by 27.20%–117.48%. It converges 30% faster than PPO from scratch (0.5M vs. 0.7M timesteps) and 20% faster than PCPO (0.62M timesteps).

2) Constraint Violation Rate: Fig. 3 presents the constraint violation rate. The proposed method significantly outperforms both baselines, achieving a better reward-safety trade-off:

- It maintains a violation rate below 1%, two orders of magnitude lower than PPO from scratch (12.3%–28.7%) and 2.14%–76.90% lower than PCPO across environments.
- PPO prioritizes reward leading to frequent violations; PCPO improves safety but relies on fixed penalties, causing over-conservatism or insufficient control. The proposed method's generalized risk network and adaptive threshold ensure safety without sacrificing reward.

These results demonstrate that the proposed method achieves a more favorable task-safety trade-off than PPO from scratch and PCPO, especially in cross-morphology transfer. It avoids PCPO's fixed penalty limitation, realizing better adaptability while maintaining superior or comparable safety.

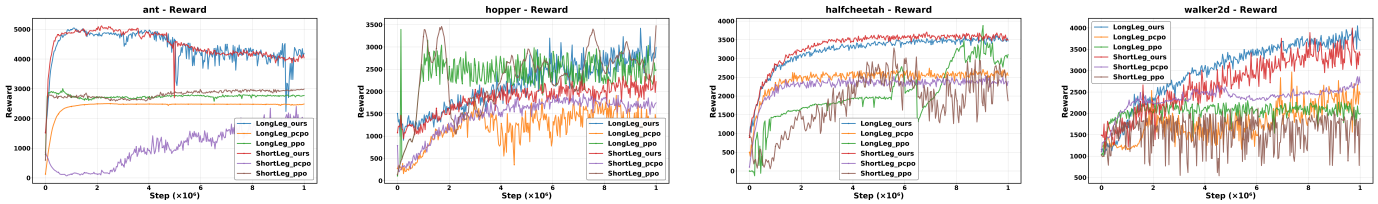
D. Ablation Study

To validate the contributions of key components in the proposed method, we conduct ablation experiments by removing individual components and evaluating the resulting variants, with the representative Ant-LongLeg environment selected as the testbed for all ablation trials. The ablation variants are defined as follows:

- **w/o generalized risk Network**: Removes the h_ψ network, using only fixed cost normalization without generalized risk mapping.
- **w/o Adaptive Threshold**: Retains the h_ψ network but uses a fixed threshold $\delta = 0.5$ instead of the adaptive threshold mechanism.
- **w/o Reference Normalization**: Initializes the policy randomly instead of using the pre-trained source policy, while retaining the generalized risk network and adaptive threshold.

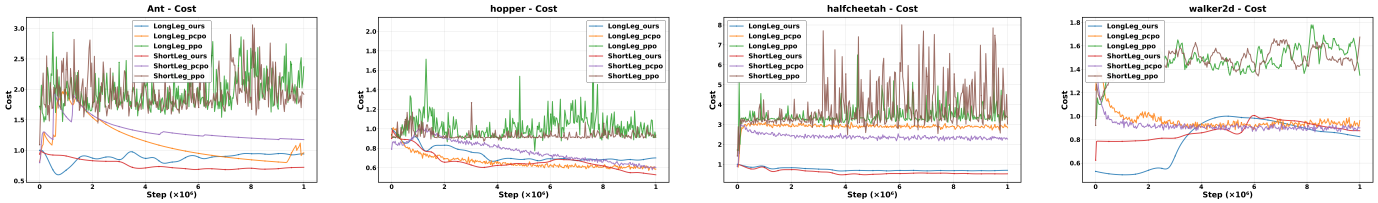
Fig. 4 shows the reward and constraint violation rate of the ablation variants. The key findings are:

1) **Importance of generalized risk Network**: The "w/o generalized risk Network" variant exhibits a constraint violation rate increase of 39.67% (from 0.712 to 0.510) and a reward drop of 49.86% compared to the full method. This confirms that the generalized risk mapping effectively captures the safety structure of the environment and guides the policy to avoid unsafe actions.
2) **Necessity of Adaptive Threshold**: The "w/o Adaptive Threshold" variant has a constraint violation rate increase of 41.27% (from 0.712 to 0.504) and a reward reduction of 49.71%. This indicates that the adaptive threshold mechanism dynamically balances performance and safety, adapting to the varying constraint requirements of different morphologies.



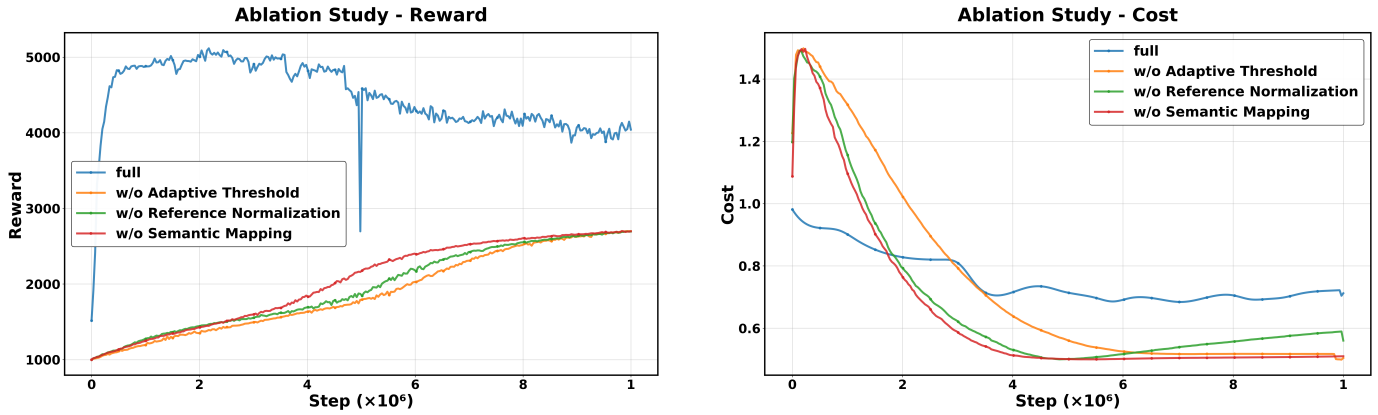
(a) Ant (b) Hopper (c) Walker2d (d) HalfCheetah

Fig. 2: Reward learning curves across environments. Comparison of our method (ours), PPO, and PCPO for both ShortLeg and LongLeg variants.



(a) Ant (b) Hopper (c) HalfCheetah (d) Walker2d

Fig. 3: Constraint satisfaction (cost) curves across environments. Lower values indicate better constraint satisfaction.



(a) Reward (b) Cost

Fig. 4: Ablation study showing the contribution of each component in our method.

3) **Value of Reference Normalization**: The "w/o Reference Normalization" variant converges much slower (requiring 1M timesteps to stabilize) and has a lower reward (49.99% reduction) compared to the full method. This highlights that reference normalization stabilizes the learning process and provides effective value estimation, accelerating adaptation to target morphologies.

These ablation results demonstrate that each component of the proposed method plays a critical role in enabling effective and safe cross-morphology policy transfer, with all three components contributing equally to reward improvement (around 50%) and reference normalization achieving the most favorable reward-cost trade-off with the smallest constraint violation increase (27.16%).

E. Discussion

The experimental results validate three key advantages of the proposed method:

1) **Zero-Shot Transfer Capability**: The method effectively adapts to morphology shifts without additional environment interaction, outperforming direct transfer in both performance and safety.
2) **Superior Trade-off Between Performance and Safety**: Compared to PPO and PCPO from scratch, the proposed method achieves higher or comparable reward while maintaining a constraint violation rate below 1%.
3) **Efficient Adaptation**: Leveraging the pre-trained source policy and generalized risk network, the method avoids inefficient exploration and converges faster than traditional reinforcement learning algorithms.

The success of the proposed method can be attributed to

two core design choices: (1) the generalized risk network preserves the safety structure of the source environment and maps it to target morphologies, and (2) the adaptive threshold mechanism dynamically adjusts the safety constraint to balance performance and safety. These design choices address the key challenges of cross-morphology transfer, namely the distribution shift caused by morphology changes and the trade-off between performance and safety.

VI. Conclusion

This paper addressed cross-morphology policy transfer from the perspective of constrained reinforcement learning. We showed that safety constraints defined in absolute cost space often fail to generalize across robot morphologies due to morphology-induced shifts in cost distributions. To mitigate this issue, we proposed a generalized-risk constrained policy optimization framework that extends PCPO with morphology-aware constraint adaptation.

The proposed method normalizes raw cost signals using reference statistics and maps them into a semantic risk space, enabling constraint enforcement that is invariant to cost scale changes. An adaptive thresholding mechanism further balances safety and performance during policy transfer. Experiments on multiple MuJoCo locomotion environments with diverse morphology variants demonstrate that our approach improves transfer stability and significantly reduces constraint violations while maintaining competitive task performance. Ablation results confirm the effectiveness of each component in the proposed framework.

Acknowledgments

We acknowledge the major support from grant (no. U2541204) .

References

[1] J. Schulman, S. Levine, P. Moritz, M. Jordan, and P. Abbeel, "Trust region policy optimization," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 1889–1897.

[2] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[3] C. Devin, A. Gupta, T. Darrell, P. Abbeel, and S. Levine, "Learning modular neural network policies for multi-task and multi-robot transfer," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, Singapore, 2017, pp. 2169–2176.

[4] T. Wang, Y. Liu, and H. Li, "Cross-embodiment robot manipulation skill transfer using latent space alignment," *arXiv preprint arXiv:2406.01968*, 2024.

[5] M. Przystupa *et al.*, "Efficient morphology-aware policy transfer to new embodiments," *Reinforcement Learning Journal*, 2025.

[6] A. Rajeswaran, S. Ghotra, S. Levine, and J. Morimoto, "EPOpt: Learning robust neural network policies using model ensembles," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Toulon, France, 2017.

[7] S. Yang, Z. Fu, Z. Cao, G. Junde, P. Wensing, W. Zhang, and H. Chen, "Multi-Loco: Unifying multi-embodiment legged locomotion via reinforcement learning augmented diffusion," in *Proc. Conf. Robot Learn. (CoRL)*, 2025, pp. 1030–1048.

[8] B. Ai, L. Dai, N. Bohlinger, D. Li, T. Mu, Z. Wu, K. Fay, H. I. Christensen, J. Peters, and H. Su, "Towards embodiment scaling laws in robot locomotion," in *Proc. Conf. Robot Learn. (CoRL)*, 2025, pp. 3483–3515.

[9] C. Devin, A. Gupta, T. Darrell, P. Abbeel, and S. Levine, "Learning modular neural network policies for multi-task and multi-robot transfer," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, Singapore, 2017, pp. 2169–2176.

[10] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Sydney, Australia, 2017, pp. 1126–1135.

[11] J. García and F. Fernández, "A comprehensive survey on safe reinforcement learning," *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 1437–1480, 2015.

[12] C. Tessler, D. J. Mankowitz, and S. Mannor, "Reward constrained policy optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, New Orleans, LA, USA, 2019.

[13] C. Schaff, C. Yunis, O. Gonzalez, A. Stern, and B. Katz, "Jointly learning to construct and control agents using deep reinforcement learning," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2019, pp. 9798–9804.

[14] A. Ray, J. Achiam, and D. Amodei, "Benchmarking safe exploration in deep reinforcement learning," *arXiv Preprint arXiv:1910.01708*, 2019.

[15] E. Altman, *Constrained Markov Decision Processes*. Boca Raton, FL, USA: CRC Press, 1999.

[16] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Sydney, Australia, 2017, pp. 22–31.

[17] L. Yang, J. Zhang, B. Li, et al., "Projection-based constrained policy optimization," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020.

[18] Y. Chow, M. Ghavamzadeh, L. Janson, and M. Pavone, "Risk-constrained reinforcement learning with percentile risk criteria," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 6070–6120, 2017.

[19] V. S. Borkar, "Risk-sensitive reinforcement learning," *IEEE Control Syst. Mag.*, vol. 32, no. 2, pp. 31–50, Apr. 2012.

[20] G. Dalal, D. Gilboa, S. Mannor, and N. Shimkin, "Safe exploration in continuous action spaces," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018.

[21] Z. Sun, S. He, F. Miao, and S. Zou, "Constrained reinforcement learning under model mismatch," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2024, pp. 47017–47032.

[22] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "A theory of learning from different domains," *Mach. Learn.*, vol. 79, no. 1–2, pp. 151–175, 2010.

[23] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv preprint arXiv:1907.02893*, 2019.

[24] M. G. Bellemare, W. Dabney, and R. Munos, "A distributional perspective on reinforcement learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 449–458.

[25] W. Dabney, M. Rowland, M. G. Bellemare, and R. Munos, "Distributional reinforcement learning with quantile regression," in *Proc. AAAI Conf. Artif. Intell.*, 2018.

[26] N. Heess et al., "Emergence of locomotion behaviours in rich environments," *arXiv preprint arXiv:1707.02286*, 2017.

[27] X. B. Peng, P. Abbeel, S. Levine, and M. van de Panne, "DeepMimic: Example-guided deep reinforcement learning of physics-based character skills," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–14, 2018.

[28] E. Todorov, T. Erez, and Y. Tassa, "MuJoCo: A physics engine for model-based control," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2012, pp. 5026–5033.

[29] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "OpenAI Gym," *arXiv Preprint arXiv:1606.01540*, 2016.