

000 001 002 003 004 005 FOCUS: HIGH-DIMENSIONAL MODEL-BASED RL VIA 006 FOCUSED CONTROL UNIT SAMPLING 007 008 009

010 **Anonymous authors**
011

012 Paper under double-blind review
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044
045
046
047
048
049
050
051
052
053

ABSTRACT

Reinforcement learning (RL) in high-dimensional continuous state and action spaces often struggles with low learning efficiency and limited exploration scalability. To address this, we introduce FOCUS, a novel model-based RL framework that leverages the insight that effective policies often rely on dynamically focused, sparse control. FOCUS learns preferences over action dimensions to facilitate more targeted and efficient policy learning. It employs a hierarchical decision-making strategy, in which a high-level policy generates binary prompts to activate control units that have more impact on the task performance, while a low-level policy produces actions conditioned on these prompts. To promote behavioral diversity guided by different control-unit preferences, we integrate a diversity-driven objective into the model-based policy optimization process. FOCUS significantly outperforms existing methods on multiple visual control tasks. **Furthermore, it facilitates the integration of prior knowledge about the importance of action dimensions, making it particularly effective for complex, high-dimensional tasks.**

1 INTRODUCTION

Model-based reinforcement learning (MBRL) provides a promising framework for decision-making tasks with high-dimensional observations by modeling environment dynamics (Hafner et al., 2020; 2025). However, its scalability is limited in high-dimensional action spaces, where exploration and policy optimization tend to become inefficient and unstable. In many real-world domains, such as humanoid locomotion, power grid control, and multi-agent collaboration, although the action spaces are high-dimensional, only a sparse subset of dimensions may be crucial for effective decision-making at each step. This insight highlights the need for mechanisms that can selectively attend to key control dimensions during learning, enabling more efficient and focused policy optimization.

A natural approach to mitigating the challenges of high-dimensional action spaces is to exploit their underlying structure. In particular, biasing policy learning toward more informative action dimensions can substantially reduce search complexity and improve efficiency. Several prior methods have explored this direction by introducing hierarchical or factorized action representations. However, many of these approaches rely on *discrete action assumptions* (Saito et al., 2024; Chen et al., 2019; Kumar et al., 2017) or *predefined base-action set* into which the original high-dimensional actions are decomposed (Kim & Dean, 2002; Geīser et al., 2020; Pierrot et al., 2021), thereby limiting their scalability to general continuous control tasks.

To address the under-explored problem of *jointly handling high-dimensional visual observations and high-dimensional continuous action spaces*, we propose FOCUS, a scalable MBRL framework. This setting presents unique challenges, including increased sample complexity and the difficulty of learning effective policies in expansive state-action spaces. Unlike existing MBRL methods (Hafner et al., 2025; Hansen et al., 2023) that treat all action dimensions uniformly, as shown in Figure 1, FOCUS introduces a *control-unit preference learning* mechanism that automatically identifies promising action subspaces and adaptively steers behavior learning toward them, enabling more efficient and focused behavior optimization and exploration.

Specifically, FOCUS represents control-unit preferences using learnable Bernoulli distributions over all action dimensions. The agent presents a *hierarchical decision-making strategy*: (i) A high-level preference policy predicts sampling probabilities and generates binary preference codes (the “prefer-

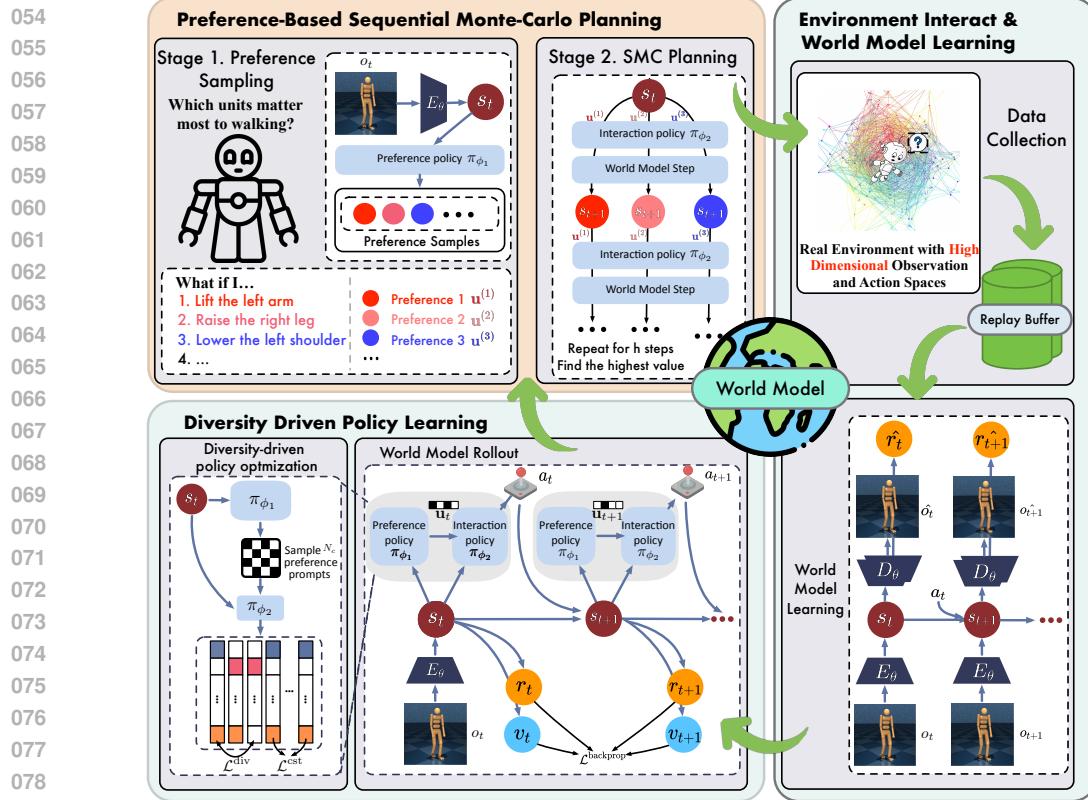


Figure 1: **FOCUS** solves high-dimensional continuous control tasks. The key idea is to improve policy optimization and exploration efficiency via a control-unit preference learning mechanism that adaptively selects and optimizes a dynamic subset of action dimensions.

ence prompt”) over all control units; (ii) a low-level interaction policy, conditioned on this prompt, specializes in the selected action subspaces. Both policies are jointly optimized to maximize expected returns over imagined state rollouts produced by the world model. To encourage model-based exploration and alignment between decision layers, we introduce a *diversity-driven policy optimization objective* that maximizes behavioral discrepancies across different control-unit preferences. Once trained, the agent leverages the hierarchical policies for real-environment interaction by performing *preference-based Monte Carlo planning*, which reduces possible control-unit sampling bias and significantly improves the decision quality.

Experiments on benchmarks with high-dimensional action spaces, *i.e.*, DeepMind Control (Tassa et al., 2018), MyoSuite (Caggiano et al., 2022), and **HumanoidBench** (Sferrazza et al., 2024), demonstrate that FOCUS outperforms model-based and model-free RL approaches in both learning efficiency and final performance. In summary, the main contributions of this work are as follows:

- We present a unified framework for *high-dimensional MBRL*. It improves existing MBRL methods for efficient learning in large action spaces. Our method dynamically prioritizes promising control units for optimization and planning by modeling the importance of individual action dimensions.
- To enable the interaction policy to respond to varying control-unit preferences, we propose *diversity-driven model-based policy learning*, which leverages a world model to imagine and evaluate trajectories under diverse preference samplings.
- As an additional contribution, we introduce a *preference-based Sequential Monte Carlo planning* algorithm, which delivers higher computational efficiency and decision quality than CEM based planning approaches in high-dimensional RL, including TD-MPC2 (Hansen et al., 2023).

2 PROBLEM FORMULATION

Our work targets high-dimensional continuous visual control tasks, where both the observation and action spaces are large and complex, posing significant challenges for efficient learning and

108 exploration. Formally, we formulate the problem as a Partially Observable Markov Decision Process
 109 (POMDP), defined by a 5-tuple $(\mathcal{O}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma)$, where: \mathcal{O} is the high-dimensional visual observation
 110 space, \mathcal{A} is the high-dimensional continuous action space, \mathcal{R} is the reward function, \mathcal{T} is the state
 111 transition dynamics, and γ is the discount factor. In particular, we assume the action space is
 112 structured as a Cartesian product of d continuous subspaces:

$$\mathcal{A} = \mathcal{A}^{(1)} \times \mathcal{A}^{(2)} \times \dots \times \mathcal{A}^{(d)}, \quad \text{with } \mathcal{A}^{(i)} \subset \mathbb{R}. \quad (1)$$

113 An action at time step t is thus represented as a d -dimensional vector $[a_t^{(1)}, a_t^{(2)}, \dots, a_t^{(d)}]$, where
 114 each $a_t^{(i)} \in \mathcal{A}^{(i)}$ is referred to as the i -th sub-action. The objective is to learn a policy $\pi: \mathcal{O} \rightarrow \mathcal{A}$,
 115 that maximizes the expected discounted cumulative return $\mathbb{E} \left[\sum_{t=1}^T \gamma^{t-1} r_t \right]$.
 116

117 **Challenges.** The entanglement of high-dimensional visual observations and high-dimensional
 118 continuous actions presents unique and under-explored challenges. Specifically, the state space
 119 induced by visual inputs (*e.g.*, raw images) is often noisy, redundant, and partially observable,
 120 requiring agents to infer latent dynamics over time. Meanwhile, large continuous action spaces lead
 121 to exponentially increasing possibilities for control decisions. Their intersection leads to an expansive
 122 and entangled state-action space, making exploration inefficient and value estimation difficult. As a
 123 result, existing RL methods often struggle with sample efficiency and generalization in such settings.
 124

125 3 METHOD

126 3.1 OVERVIEW OF FOCUS

127 To tackle the challenges described above, we propose an MBRL method guided by learned control-
 128 unit preferences. The overall training and environment interaction pipeline consists of three stages:

- 129 (i) *World model learning*: We first train a world model to predict latent state transitions and reward
 130 signals from observation-action-reward tuples. We use DreamerV3 (Hafner et al., 2025) as the
 131 backbone for world modeling. Detailed model specifications are provided in Appendix A
- 132 (ii) *Diversity-driven model-based policy learning*: As shown in Algorithm 1, over future trajectory
 133 rollouts (*i.e.*, latent state imaginations) generated by the world model, FOCUS constructs a
 134 hierarchical policy learning architecture. At its core is a high-level decision-making module
 135 that outputs a preference distribution over all control units¹. At each imagination step, the
 136 high-level policy samples a binary control-unit preference from Bernoulli distributions (termed
 137 a “*preference prompt*”). This prompt selects a sparse subset of control units, guiding the
 138 low-level policy network to generate specific behaviors within the original high-dimensional
 139 action space. The high-level *preference policy* and the low-level *interaction policy* are jointly
 140 optimized to maximize expected returns over the imagination horizon. Furthermore, we
 141 encourage behavioral diversity across different preference prompts, promoting more effective
 142 exploration and stronger alignment between hierarchical decisions.
 143
- 144 (iii) *Preference-based Monte Carlo planning*: During environment interaction, the agent samples
 145 multiple candidate preference prompts. For each prompt, it performs multi-step Monte Carlo
 146 planning to generate potential future trajectories. The final action is then selected from these
 147 candidates based on trajectory weights computed from their cumulative advantage estimates.
 148

149 3.2 CONTROL-UNIT PREFERENCE SAMPLING

150 Our approach is motivated by the observation that high-dimensional control tasks often exhibit inherent
 151 sparsity—at any given time, some control units more significantly contribute to task performance
 152 than others. For example, in the HumanoidStand task, the agent begins by pushing itself upward
 153 from the ground using support from its legs and arms. In such moments, joints such as the ankles
 154 and wrists play a dominant role, while many other action dimensions contribute minimally. This
 155 observation suggests that exploration can be made more efficient by selectively attending to a small
 156 subset of all control units. Accordingly, we introduce a high-level policy π_{ϕ_1} that takes latent states as
 157 inputs and generates a binary prompt $\mathbf{u}_t \in \mathbb{R}^d$, where $\mathbf{u}_t^{(i)} = 1$ indicates that the i -th action dimension

158 ¹Each *control unit* corresponds to a subset of action dimensions for a specific functional component (*e.g.*, a
 159 joint in a humanoid robot), potentially spanning multiple dimensions to reflect its degrees of freedom.

162 **Algorithm 1:** Full training algorithm of FOCUS, with **Diversity-driven model-based policy**
163 **learning**, and **Preference-based SMC Planning**

1 **Initialize:** world model θ , hierarchical actor $\{\phi_1, \phi_2\}$; critic ψ , replay buffer \mathcal{B} with random episodes.

2 **while** not converged **do**

3 **for** update step $c = 1 \dots C$ **do**

4 // World model learning

5 Draw data sequences $\{(o_t, a_t, r_t)\}_{t=1}^T \sim \mathcal{B}$.

6 Update the dynamics model p_θ and reward model r_θ using DreamerV3 world model objectives.

7 // Diversity-driven model-based policy learning

8 Draw a random observation $o_1 \sim \mathcal{B}$.

9 **for** time step $t = 1 \dots H - 1$ **do**

10 Sample N_c control-unit preference prompts $\{\mathbf{u}_{t,n}\}_{n=1}^{N_c} \sim \pi_{\phi_1}(s_t)$.

11 Generate $\{a_{t,n} \sim \pi_{\phi_2}(s_t, \mathbf{u}_{t,n})\}_{n=1}^{N_c}$.

12 Predict $\{s_{t+1,n} \sim p_\theta(s_t, a_{t,n})\}_{n=1}^{N_c}$ and $\{r_{t,n} \sim r_\theta(s_{t+1,n})\}_{n=1}^{N_c}$.

13 Compute $\mathcal{L}_t^{\text{div}}$ and $\mathcal{L}_t^{\text{est}}$ over $\{a_{t,n}\}_{n=1}^{N_c}$ according to Eq. equation 3 and Eq. equation 4.

14 Select $s_{t+1,n}$ with the highest $v_\psi(s_{t+1,n})$ as s_{t+1} .

15 **end**

16 Update π_{ϕ_1} , π_{ϕ_2} and v_ψ using Eq. equation 5.

17 **end**

18 // Environment interaction with sequential Monte Carlo planning

19 $o_1 \leftarrow \text{env.reset}()$

20 **for** time step $t = 1 \dots T$ **do**

21 Compute posterior state $s_t \sim q_\theta(o_t)$.

22 Sample N_p preference prompts $\{\mathbf{u}_n\}_{n=1}^{N_p} \sim \pi_{\phi_1}(\cdot | s_t)$.

23 Initialize N_p planning trajectories $\{\tau_n\}_{n=1}^{N_p}$ starting from $\{\hat{s}_{t,n}\}_{n=1}^{N_p} = s_t$.

24 Initialize importance weights for planning trajectories $\{w_n = 1\}_{n=1}^{N_p}$.

25 **for** planning steps $h = 0 \dots L - 1$ **do**

26 Generate $\{\hat{a}_{t+h,n} \sim \pi_{\phi_2}(\hat{s}_{t+h,n}, \mathbf{u}_n)\}_{n=1}^{N_p}$.

27 Predict $\{\hat{s}_{t+h+1,n} \sim p_\theta(\hat{s}_{t+h,n}, \hat{a}_{t+h,n})\}_{n=1}^{N_p}$ and $\{\hat{r}_{t+h,n} \sim p_\theta(\hat{s}_{t+h+1,n})\}_{n=1}^{N_p}$.

28 Compute step-wise advantages $\{A_{t+h,n} = \hat{r}_{t+h,n} + \gamma v_\psi(\hat{s}_{t+h+1,n}) - v_\psi(\hat{s}_{t+h,n})\}_{n=1}^{N_p}$.

29 Update weights $\{w_{t+h+1,n} = w_{t+h,n} \cdot \exp(A_{t+h,n})\}_{n=1}^{N_p}$.

30 **end**

31 Sample τ_n from a categorical distribution based on weights $\{w_{t+L,n}\}_{n=1}^{N_p}$.

32 Select the first action $\hat{a}_{t,n}$ in τ_n as a_t .

33 $r_t, o_{t+1} \leftarrow \text{env.step}(a_t)$

34 **end**

35 Add experience to the replay buffer $\mathcal{B} \leftarrow \mathcal{B} \cup \{(o_t, a_t, r_t)\}_{t=1}^T$.

36 **end**

is currently important and should be prioritized. This design leads to a hierarchical policy structure, where the target policy π_{ϕ_2} is conditioned on both the latent state and the preference prompt:

$$\text{Preference policy: } \mathbf{u}_t \sim \pi_{\phi_1}(s_t), \quad \text{Interaction policy: } a_t \sim \pi_{\phi_2}(\mathbf{u}_t, s_t), \quad (2)$$

where \mathbf{u}_t is sampled from d independent Bernoulli distributions whose parameters are jointly predicted by the high-level policy π_{ϕ_1} . Despite the independent sampling process, co-optimizing the high-level and low-level policies implicitly facilitates the discovery of interdependencies among action dimensions. For comparisons between independent sampled Bernoulli variables and explicitly enforced inter-unit dependencies, please refer to Appendix C.7. Intuitively, we expect the target policy π_{ϕ_2} to adapt its behavior based on the high-level guidance \mathbf{u}_t , enabling more focused and effective exploration within the subspace of prioritized action dimensions.

3.3 DIVERSITY-DRIVEN MODEL-BASED POLICY LEARNING

To promote an efficient exploration of potential policies guided by diverse control-unit activations, we introduce a focused policy optimization approach within a model-based behavior learning framework, as shown in Figure 1. We first learn a world model parameterized by θ , which includes an encoder $p_\theta(s_t | o_t)$, a latent transition model $p_\theta(s_{t+1} | s_t, a_t)$, and a reward predictor $r_\theta(s_t)$. Given an initial

latent state encoded from a random observation in the replay buffer, FOCUS samples a preference prompt \mathbf{u}_t , and generates an preference-conditioned action a_t based on the current rollout state \hat{s}_t . The action a_t is then used to update the imagination trajectory via the world model. The hierarchical policy $\pi_{\phi_1,2}$ and the value model v_ψ are optimized over imagined trajectories of length H .

However, directly maximizing the expected value function, as in Dreamer-style objectives, is insufficient for inducing low-level behaviors that are responsive to high-level preferences. Without further regularization, the actor may fail to interpret and follow the high-level guidance effectively. To address this, we design a diversity-driven learning objective. The key idea is that the low-level actor should produce diverse behaviors for control units selected by different preference prompts, while maintaining consistent behaviors for those not selected. Specifically, we sample N_c preference prompts $\{\mathbf{u}_{t,n}\}_{n=1}^{N_c}$ from $\pi_{\phi_1}(s_t)$, and generate corresponding action vectors $\{a_{t,n}\}_{n=1}^{N_c}$. Each action is decomposed as $a_{t,n} = [a_{t,n}^{(1)}, \dots, a_{t,n}^{(d)}]$ along d dimensions. We first define a diversity-promoting loss to encourage the low-level policy to respond distinctly to different preference selections:

$$\mathcal{L}_t^{\text{div}} = \sum_{i=1}^d \left[- \sum_{n=1}^{N_c} \mathbb{I}(\mathbf{u}_n^{(i)} = 1) \sum_{k=1}^{N_c} \mathbb{I}(\mathbf{u}_k^{(i)} = 0) \cdot \text{KL} \left[\pi(a_n^{(i)} | s, \mathbf{u}_n) \parallel \text{sg}(\pi(a_k^{(i)} | s, \mathbf{u}_k)) \right] \right], \quad (3)$$

where $\mathbb{I}(\cdot)$ is the indicator function, $\text{sg}(\cdot)$ denotes stop-gradient, and we ignore the time step index for clarity. Meanwhile, we introduce a consistency loss to penalize divergence among actions on the same dimensions not selected by the preference prompt:

$$\mathcal{L}_t^{\text{cst}} = \sum_{j=1}^d \left[\sum_{n=1}^{N_c} \mathbb{I}(\mathbf{u}_n^{(j)} = 0) \sum_{k=1}^{N_c} \mathbb{I}(\mathbf{u}_k^{(j)} = 0) \cdot \text{KL} \left[p(a_n^{(j)} | s, \mathbf{u}_n) \parallel \text{sg}(p(a_k^{(j)} | s, \mathbf{u}_k)) \right] \right]. \quad (4)$$

Together, these losses impose a contrastive structure in the high-dimensional action space, enhancing exploration and ensuring better alignment between low- and high-level policies. Importantly, we do not stop the gradient at \mathbf{u}_n in the above regularization terms. Instead, we allow π_{ϕ_1} to be co-optimized with π_{ϕ_2} . This design further promotes exploration, as $\mathcal{L}_t^{\text{div}}$ rewards diverse control-unit activations and prevents the high-level policy from collapsing to identical but biased activation vectors.

For the sake of computational efficiency, as shown in Algorithm 1, we choose $s_{t+1,n} \sim p_\theta(s_t, a_{t,n})$ with the highest predicted value $v_\psi(s_{t+1,n})$ as s_{t+1} among multiple candidates at each imagination step. In practice, we compute the diversity-driven objectives only at the first step of imagination. Despite this simplification, we observe a significant performance gain over versions without diversity-driven regularization, while incurring almost no additional training overhead.

Finally, the full objective is defined as follows, where the actor uses value backpropagation through dynamics as in DreamerV3 for better continuous control, and the critic uses a λ -return target:

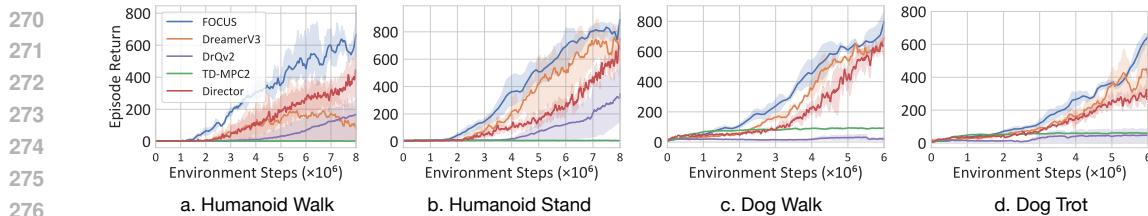
$$\begin{aligned} \text{Actor: } \mathcal{L}(\phi_1, \phi_2) &= \mathbb{E}_{p_\psi, p_\theta} \left[\sum_{t=1}^{H-1} \left[\mathcal{L}_t^{\text{div}} + \mathcal{L}_t^{\text{cst}} - V_t^\lambda - \eta \mathcal{H}(a_t | \hat{s}_t, \text{sg}(\mathbf{u}_t)) \right] \right], \\ \text{Critic: } \mathcal{L}(\psi) &= \mathbb{E}_{p_{\phi_1}, p_{\phi_2}, p_\theta} \left[\sum_{t=1}^{H-1} \frac{1}{2} (v_\psi(\hat{s}_t) - \text{sg}(V_t^\lambda))^2 \right], \end{aligned} \quad (5)$$

where the entropy term $\mathcal{H}(\cdot)$ further encourages exploration conditioned on preference prompts. Notably, both levels of policies are jointly optimized towards maximizing the imagined values.

3.4 PREFERENCE-BASED MONTE CARLO PLANNING

The hierarchical policy involves two sources of stochasticity—from both the high-level preference sampling and the low-level action generation—which may increase variance during real-environment interaction. To mitigate this, we develop an online planning scheme that efficiently reasons over multiple future state-action trajectories, each conditioned on a different control-unit preference prompt. We term this method *preference-based Monte Carlo planning*, a modified form of sequential Monte Carlo planning (Piché et al., 2018) designed for high-dimensional action spaces.

As shown in Algorithm 1, at each interaction step, we sample N_p preference prompts $\{\mathbf{u}_n\}_{n=1}^{N_p}$ based on the current state s_t , and initialize N_p planning trajectories with uniform weights $w_n = 1$. At

Figure 2: **Results on DeepMind Control Suite.** All results are averaged over three training seeds.

each planning step $t + h$, we reuse \mathbf{u}_n to sample low-level actions $\hat{a}_{t+h,n} \sim \pi_{\phi_2}(\hat{s}_{t+h,n}, \mathbf{u}_n)$. This contrasts with model-based policy learning, where \mathbf{u}_n is typically re-sampled at each imagination step. Instead, we preserve temporal consistency in the sampled control-unit prompts during planning. We then use the learned dynamics p_θ to rollout future states $s_{t+h:t+L,n}$. We evaluate these candidate trajectories via advantage estimation and update their weights accordingly:

$$A_{t+h,n} = \hat{r}_{t+h,n} + \gamma v_\psi(\hat{s}_{t+h+1,n}) - v_\psi(\hat{s}_{t+h,n}), \quad w_{t+h+1,n} = w_{t+h,n} \cdot \exp(A_{t+h,n}). \quad (6)$$

At the end of the planning horizon, we sample one trajectory from the categorical distribution defined by $\{w_{t+L,n}\}_{n=1}^{N_p}$ and execute its first action in the environment. Empirically, this approach significantly outperforms prior sample-based planning methods (*e.g.*, MPPI in TD-MPC2 (Hansen et al., 2023)) in both computational efficiency and decision quality. Unlike MPPI, our method avoids online policy refitting, making it more suitable for high-dimensional control tasks.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

First, we conduct experiments on two RL environments with visual inputs, namely DeepMind Control Suite (Tassa et al., 2018) and MyoSuite (Caggiano et al., 2022). We present empirical results on seven tasks from these environments with high-dimensional continuous action spaces. **To demonstrate the applicability of our diversity-driven hierarchical policy learning, we conduct additional experiments on two tasks of HumanoidBench (Sferrazza et al., 2024) with proprioceptive robot state as input.** Detailed information on the selected tasks with their action space sizes is provided in Appendix B.

We compare FOCUS with: (1) *DreamerV3* (Hafner et al., 2025), a strong model-based RL method in high-dimensional observation space; (2) *TD-MPC2* (Hansen et al., 2023), a decoder-free model-based RL baseline that employs the CEM-based planning method; (3) *DrQv2* (Yarats et al., 2021a), a data augmentation-based visual RL approach; (4) *Director* (Hafner et al., 2022), another hierarchical MBRL method that conditions low-level policy on goals generated by the high-level autoencoder. **Particularly for HumanoidBench, the previous model-based methods suffer from computational overhead.** While FastTD3 (Seo et al., 2025) has demonstrated superior efficiency in solving HumanoidBench tasks, we make modifications to this model-free baseline to create a new variant named FastTD3-FOCUS, and compare it to the vanilla FastTD3. The modified version substitutes hierarchical policy for a flat one, incorporating diversity-driven learning objective in FOCUS, without learning a world model and execute SMC planning based on the learnt model.

All models are trained under a fixed environment interaction budget. For two tasks in the Humanoid suite, each model is trained for 8M frames, and for the Dog suite, 6M frames. **Tasks in MyoSuite allow interactions for 1M frames, and 500K for tasks in HumanoidBench.** Average episode returns are computed over 10 episodes per seed.

4.2 DOMAIN #1: DEEPMIND CONTROL SUITE

As shown in Figure 2, FOCUS consistently outperforms baseline methods across all four tasks. Notably, it achieves a substantial improvement over DreamerV3 in HumanoidWalk. In both Humanoid tasks, the learning curves show a prolonged warm-up phase before meaningful improvement, reflecting the challenge of locating sparse high-reward regions within large decision spaces. Compared with **DreamerV3**, our model exhibits an earlier performance rise, demonstrating its ability to reduce sample complexity and efficiently explore targeted reward regions. In Dog tasks, rewards are less sparse than in Humanoid tasks, and DreamerV3 is already capable of sampling high-reward regions. This is reflected in its steadily increasing learning curves, suggesting that despite the larger

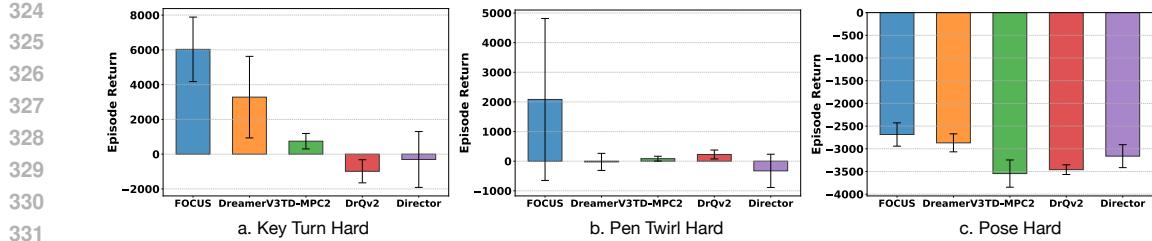


Figure 3: **Results on the MyoSuite hard tasks.** Episode returns are averaged over 10 episodes per seed, for a total of 30 runs. FOCUS outperforms all baselines across every task.

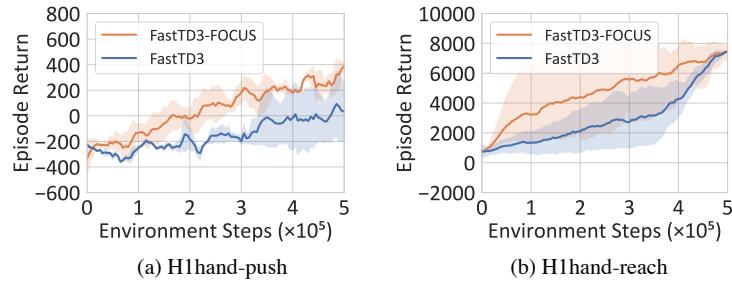


Figure 4: **Results on HumanoidBench.** We implement FOCUS on top of FastTD3 rather than DreamerV3 or TD-MPC2, as these models fail to converge within the same number of interaction steps, likely due to limited interaction efficiency. FastTD3, in contrast, is specifically optimized for efficient environment interaction. All results are averaged over three training seeds.

number of action dimensions, the high-reward regions in the action space are sufficiently large for agents with random policies to sample with reasonable probability. In this setting, exploration is not the primary bottleneck, which limits potential performance gains. For **TD-MPC2**, it fails to achieve meaningful progress on humanoid tasks, highlighting its difficulty in effectively exploring the vast search space, particularly under high-dimensional visual input. The comparison with **Director** highlights the impact of the diversity-driven objective: while both methods employ hierarchical policies, FOCUS explicitly prioritizes subsets of the action space, resulting in higher exploration efficiency. We showcase the policy evaluation of FOCUS, DreamerV3 and Director in ??, with full results in Appendix C.1. As illustrated, the policy obtained by FOCUS adjusts humanoid’s posture in fewer time steps to start walking ($t = 50$), meanwhile maintaining its control more stable than DreamerV3 which falls to the ground ($t = 200$) after standing up. Additionally, DrQv2 requires roughly 4M environment steps before its learning curve begins to rise, suggesting that augmenting failing episodes early has little effect on sample efficiency.

We provide further results, comparing the planning efficiency of FOCUS with TD-MPC2 in Appendix C.2, and evaluating its generalization to low-dimensional action spaces in Appendix C.4.

4.3 DOMAIN #2: MYOSUITE

Performance results on MyoSuite are shown in Figure 3. FOCUS outperforms all baselines on KeyTurnHard and PenTwirlHard. On KeyTurnHard, it improves over DreamerV3, demonstrating higher efficiency in exploring high-reward regions within the same environment step budget. On PenTwirlHard, FOCUS successfully discovers sharply peaked reward regions, whereas other baselines struggle in flatter areas; DreamerV3 fails to solve this task. The integration of the diversity-promoting loss \mathcal{L}_t^{div} and consistency loss \mathcal{L}_t^{cst} systematically biases sampling toward high-reward regions, enabling efficient policy learning and sustained performance gains. All methods face challenges on PoseHard, though FOCUS achieves slightly better results, thanks in part to \mathcal{L}_t^{cst} preventing overfitting as discussed in Section 3.3.

4.4 DOMAIN #3: HUMANOIDBENCH

We present the results of two tasks on HumanoidBench in Figure 4, while results of extensive 6 tasks could be found in Figure 16. We implement FOCUS on top of FastTD3 rather than DreamerV3 or TD-MPC2, as these models fail to converge within the same number of interaction steps, likely due to limited interaction efficiency. FastTD3, in contrast, is specifically optimized for efficient

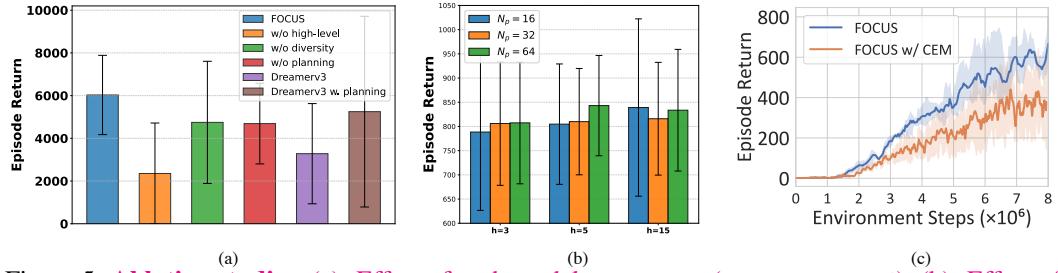


Figure 5: **Ablation studies.** (a): Effect of each model component (KeyTurnHard). (b): Effect of planning horizon h and candidate trajectories N_p in our planner (HumanoidWalk). (c): Ablation study comparing CEM and SMC planning (HumanoidWalk).

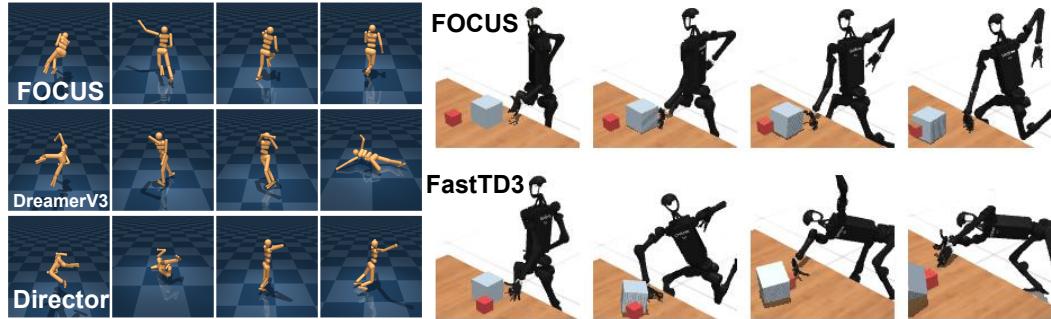


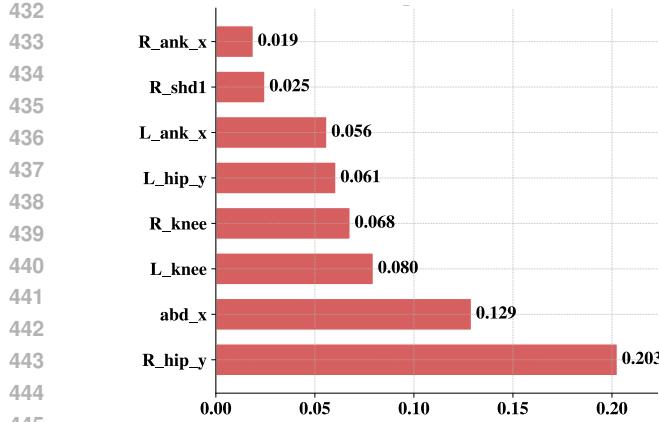
Figure 6: **Visualization of the learned policies.** We compare different models on HumanoidWalk (Left) and H1hand-push from HumanoidBench (Right). In H1hand-push, the robot’s task is to push the object to the red target location.

environment interaction. Details of FastTD3 and our modifications are provided in Appendix C.5. In both H1hand-push and H1hand-reach, FastTD3-FOCUS achieves better sample efficiency than the vanilla FastTD3, which uses a simple MLP policy. As shown in Figure 4a for H1hand-push, FastTD3-FOCUS maintains consistent improvement throughout training, indicating that focused exploration helps the agent avoid local optima where the flat MLP policy is prone to getting stuck. Similarly, in the H1hand-reach task (Figure 4b), although both methods eventually converge to similar episodic returns, FastTD3-FOCUS demonstrates higher sample efficiency. This follows from the diversity-driven objective, which encourages distinct and meaningful behavior patterns among control units, enabling faster coverage of high-return regions of the state space and accelerating the acquisition of effective reaching behaviors.

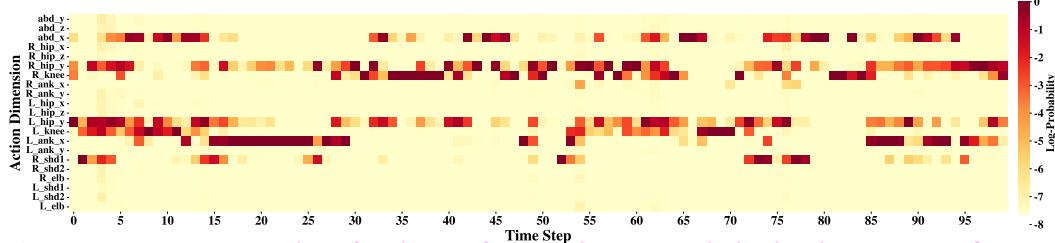
4.5 MODEL ANALYSES

Ablation studies. We validate the proposed modules on KeyTurnHard in MyoSuite, considering three variants of our method: (1) FOCUS without π_{ϕ_1} , **reducing it to a DreamerV3 agent planning in the original high-dimensional action space**; (2) FOCUS without \mathcal{L}_t^{div} and \mathcal{L}_t^{cst} ; and (3) FOCUS without planning. These variants are compared with the full FOCUS, DreamerV3, and DreamerV3 w/ planning to assess the contribution of each component. As shown in Figure 5(a), FOCUS consistently outperforms all other variants, demonstrating that each component contributes to its success. The variant without the high-level policy π_{ϕ_1} is equivalent to DreamerV3 with the addition of a Sequential Monte Carlo planner during **both exploration and policy execution**, yet it performs worse than DreamerV3 on account of biased exploration under noised value guidance. DreamerV3 with planning shows improved but more unstable behavior, indicating vulnerability to being trapped in local optima. Instead of planning in the flat action space, FOCUS plans in high level which **promotes diversity, effectively bypassing these optimization traps during training**. Furthermore, both the diversity loss and the planning algorithm are crucial contributors to FOCUS’s performance. We also compare FOCUS to its Cross Entropy Method (CEM)-based variant in Figure 5(c). As illustrated, SMC enables faster and more effective learning compared to CEM. We argue that prior sample-based planning methods like CEM, which require online distribution refitting, are not sufficiently capable of modeling target distributions in a high-dimensional action space and thus limit their effectiveness.

Hyperparameter analyses. We examine different hyperparameter combinations for planning horizons h and the number of candidate trajectories N_p in Figure 5 (b) and ???. We find that increasing



445
446
447
Figure 7: **Visualization of most frequently selected action dimensions.** The sparse action dimensions highlighted by FOCUS align closely with the kinematic priorities of the HumanoidStand task, demonstrating that the model effectively identifies behaviorally relevant control channels.



455
456
457
Figure 8: **Temporal evolution of action preferences in Bernoulli distributions over the first 100 timesteps.** The high-level preference policy employs a sparse and dynamic selection strategy, with specific action dimensions being selectively activated in response to temporal transitions.

459
460
461
462
463
464
465
 N_p generally improves overall performance, but does not guarantee a monotonic increase in episode return. We attribute this to the exclusive use of the policy log-likelihood in the advantage calculation, which can make planning riskier. Including the log-likelihood term is standard practice in previous Control-as-Inference research (Levine, 2018; Piché et al., 2018), but in high-dimensional action spaces it introduces large variance, leading to rapid particle degeneration. Removing this term mitigates variance but promotes riskier planning, as actions with larger magnitudes are weighted higher without accounting for the probability of being sampled.

4.6 VISUALIZATION OF LEARNED CONTROL-UNIT PREFERENCES

466
467
468
We further present quantitative and qualitative results regarding the preference prompts sampled on HumanoidStand, using the model trained immediately after the rising point (*i.e.*, after 3M environment steps) and recording preference prompts over a single episode. These results serve two purposes: (1) to qualitatively assess whether the learned preference policy captures meaningful, structured sparsity across action dimensions, and (2) to illustrate the temporal evolution of control preferences as the agent attempts to stand.

474
475
476
477
478
479
480
481
482
483
We visualize the 8 most frequently selected action dimensions on the humanoid embodiment along with their selection frequencies in Figure 7. The full action dimension frequencies are presented in Appendix C.8, demonstrating that FOCUS successfully captures the structural sparsity within the original action space. Actuators corresponding to these preferred dimensions are highlighted to showcase FOCUS’s ability to discover meaningful kinematic correspondences. As illustrated, the learned preferences are strongly concentrated on the lower-body actuators and the core trunk, which are physically critical for the HumanoidStand task. Specifically, the hip joints (*e.g.*, R_hip_y) and abdomen (abd_x) dominate the selection, reflecting their key role in generating upward torque and maintaining core stability necessary for an upright posture. This distribution confirms that FOCUS effectively identifies structural sparsity, facilitating efficient policy learning in a reduced action space.

484
485
In Figure 8, we further visualize the temporal evolution of the learned preference prompts in Bernoulli distributions. The results show that rather than converging to a static set of actuators, the policy dynamically modulates its focus across different dimensions over time, adapting to time-varying task

486 requirements. Together, these results support the hypothesis that FOCUS leverages structured sparsity
 487 to concentrate exploration on behaviorally relevant subspaces, aligning with our goal of reducing
 488 exploration overhead in high-dimensional control tasks.
 489

490 5 RELATED WORK 491

492 **Model-based RL.** Model-based RL (Ha & Schmidhuber, 2018; Hafner et al., 2025; Hansen et al.,
 493 2023; Moerland et al., 2023) shows great potential in addressing the data efficiency issue compared
 494 to model-free methods (Yarats et al., 2021b; Kostrikov et al., 2021; Laskin et al., 2020), particularly
 495 in online learning. By leveraging a learned world model, these methods enable efficient policy
 496 optimization through planning (Nguyen et al., 2021; Zhao et al., 2021; Hansen et al., 2023) or Q-
 497 learning (Hafner et al., 2020; Wang et al., 2022). However, in high-dimensional observation and action
 498 spaces, world models struggle to identify sparse high-reward regions due to their uniform treatment
 499 of all action dimensions. This limitation burdens the learning process, as undirected exploration in
 500 vast action spaces leads to inefficiency. FOCUS addresses this challenge by adaptively selecting a
 501 control unit, thereby guiding policy exploration and optimization within a reduced subspace.
 502

503 **Hierarchical RL.** Prior studies in hierarchical structured action space (Kumar et al., 2017; Chen
 504 et al., 2019; Tang & Agrawal, 2020; Saito et al., 2024) usually assume a multi-stage policy that divides
 505 the searching space into smaller parts. However, this structure is mostly restricted to tasks with
 506 discrete action spaces and requires additional discretization when confronted with continuous ones.
 507 Hierarchical policy learning in behavior learning (Pateria et al., 2021) has been used to enable efficient
 508 long-horizon reasoning by generating temporal abstractions through high-level policy (Vezhnevets
 509 et al., 2017; Gumsch et al., 2024; Gürler et al., 2021). Director (Hafner et al., 2022) is a model-based
 510 method featuring a hierarchical policy structure that generates a goal with a high-level autoencoder
 511 first. Yet in our work, we have made use of the hierarchical structure to model the sparsity nature
 512 of the high-dimensional action space, thus reducing the searching space to a single control unit and
 513 guiding the exploration in a high-dimensional continuous action space.
 514

515 **RL with high-dimensional action space.** Despite the hierarchical structure mentioned above,
 516 methods that exploit prior knowledge to reduce exploration can efficiently deal with high-dimensional
 517 action space. Factored action space (Osband & Van Roy, 2014; Guestrin et al., 2001; Mahajan
 518 et al., 2021; Peng et al., 2021) decompose the original MDP into several sub-questions. Other works
 519 focused on over-actuated system and leverage the prior of musculoskeletal systems to reduce action
 520 space (Schumacher et al., 2022; Luo et al., 2023). Leaning in high-dimensional action space becomes
 521 difficult when prior knowledge is inaccessible. Similar to FOCUS, it's intuitive to learn from data the
 522 redundancy within action space.(Chandak et al., 2019; Baram et al., 2021; Zhong et al., 2024), while
 523 these methods either suffer from generalization or be specific to discrete action spaces. Additionally,
 524 learning-based structure discovery methods (Tavakoli et al., 2017; 2018; Van de Wiele et al., 2020;
 525 Zhang & Cai) are trained to discover underlying structure with separate actors or deterministic action
 526 mask. However, these designs either ignore action dimensional interdependencies, or suffer from
 527 drastic-shifting dynamics, which can lead to instability during training.
 528

529 6 CONCLUSIONS AND LIMITATIONS 530

531 In this paper, we introduced FOCUS, a novel approach for high-dimensional reinforcement learning
 532 that addresses the challenges of learning and exploration efficiency in large action spaces. FOCUS
 533 leverages a hierarchical policy structure, where a high-level policy learns to generate sparse control-
 534 unit activations, identifying key action dimensions at each decision point; while a low-level policy
 535 operates within these preferred subspaces to optimize behavior using imagined rollouts from a
 536 world model. Our empirical results showed that FOCUS consistently outperforms baselines on
 537 high-dimensional control tasks while maintaining high learning efficiency.
 538

539 Despite its strong performance, FOCUS also has certain limitations. First, the effectiveness of the
 540 learned preference prompts is closely tied to the quality of the world model, as inaccuracies in imag-
 541 ined rollouts can lead to suboptimal high-level guidance. Second, the hierarchical decision-making
 542 mechanism introduces extra parameters and training complexity. Exploring effective preference
 543 representations can be a promising direction for future work.
 544

540
541 ETHICS STATEMENT542
543 This work proposes an RL framework focusing on dealing with high-dimensional action spaces. The
544 research does not involve human subjects, sensitive personal data, or any newly collected datasets.
545 All experiments were conducted within simulated environments, ensuring that no privacy or security
546 concerns were introduced.547 Our study does not directly present harmful applications. The contributions are methodological
548 and algorithmic in nature, aiming to deepen the understanding of efficient policy learning in high-
549 dimensional spaces. We have carefully read and followed the Code of Ethics, and are committed to
550 principles of research integrity. No conflicts of interest or external sponsorship influenced this work.551
552 REPRODUCIBILITY STATEMENT553
554 This work proposes a novel model, FOCUS, to deal with RL in tasks with high-dimensional observa-
555 tion and action input. To facilitate reproducibility, we provide the complete implementation code as
556 part of the supplementary materials. Detailed descriptions of the model is included in Section 3.2,
557 Section 3.3 and the planning algorithm is included in Section 3.4. We also describe the detailed
558 experimental settings in Section 4.1 and Appendix B.559
560 REFERENCES561 Nir Baram, Guy Tennenholz, and Shie Mannor. Action redundancy in reinforcement learning. In
562 Uncertainty in Artificial Intelligence, pp. 376–385. PMLR, 2021.563 Vittorio Caggiano, Huawei Wang, Guillaume Durandau, Massimo Sartori, and Vikash Kumar.
564 Myosuite—a contact-rich simulation suite for musculoskeletal motor control. arXiv preprint
565 arXiv:2205.13600, 2022.566 Yash Chandak, Georgios Theocharous, James Kostas, Scott Jordan, and Philip Thomas. Learning
567 action representations for reinforcement learning. In International conference on machine learning,
568 pp. 941–950. PMLR, 2019.569 Haokun Chen, Xinyi Dai, Han Cai, Weinan Zhang, Xuejian Wang, Ruiming Tang, Yuzhou Zhang,
570 and Yong Yu. Large-scale interactive recommendation with tree-structured policy gradient. In
571 Proceedings of the AAAI conference on artificial intelligence, volume 33, pp. 3312–3320, 2019.572 Florian Geißer, David Speck, and Thomas Keller. Trial-based heuristic tree search for mdps with
573 factored action spaces. In Proceedings of the International Symposium on Combinatorial Search,
574 volume 11, pp. 38–47, 2020.575 Carlos Guestrin, Daphne Koller, and Ronald Parr. Multiagent planning with factored mdps. Advances
576 in neural information processing systems, 14, 2001.577 Christian Gumbsch, Noor Sajid, Georg Martius, and Martin V. Butz. Learning hierarchical world mod-
578 els with adaptive temporal abstractions from discrete latent dynamics. In The Twelfth International
579 Conference on Learning Representations, 2024. URL <https://openreview.net/forum?id=TjCDNssXKU>.580 Nico Gürtler, Dieter Büchler, and Georg Martius. Hierarchical reinforcement learning with timed
581 subgoals. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.),
582 Advances in Neural Information Processing Systems, volume 34, pp. 21732–21743. Curran Asso-
583 ciates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/b59c21a078fde074a6750e91ed19fb21-Paper.pdf.584 David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In Proceedings
585 of the 32nd International Conference on Neural Information Processing Systems, NIPS’18, pp.
586 2455–2467, Red Hook, NY, USA, 2018. Curran Associates Inc.587 Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning
588 behaviors by latent imagination. In ICLR, 2020.

594 Danijar Hafner, Kuang-Huei Lee, Ian Fischer, and Pieter Abbeel. Deep hierarchical planning from
 595 pixels. *Advances in Neural Information Processing Systems*, 35:26091–26104, 2022.
 596

597 Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains
 598 through world models. *Nature*, 640:647–653, 2025.
 599

600 Nicklas Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous
 601 control. *arXiv preprint arXiv:2310.16828*, 2023.
 602

603 Kee-Eung Kim and Thomas Dean. Solving factored mdps with large action space using algebraic
 604 decision diagrams. In *Pacific Rim International Conference on Artificial Intelligence*, pp. 80–89.
 605 Springer, 2002.
 606

607 Ilya Kostrikov, Denis Yarats, and Rob Fergus. Image augmentation is all you need: Regularizing
 608 deep reinforcement learning from pixels. In *ICLR*, 2021.
 609

610 A Kumar, K Swersky, and G Hinton. Feudal learning for large discrete action spaces with recursive
 611 substructure. In *Proceedings of the NIPS Workshop Hierarchical Reinforcement Learning, Long*
 612 *Beach, CA, USA*, volume 9, 2017.
 613

614 Michael Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, and Aravind Srinivas.
 615 Reinforcement learning with augmented data. In *NeurIPS*, 2020.
 616

617 Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review.
 618 *arXiv preprint arXiv:1805.00909*, 2018.
 619

620 Shuzhen Luo, Ghaith Androwis, Sergei Adamovich, Erick Nunez, Hao Su, and Xianlian Zhou.
 621 Robust walking control of a lower limb rehabilitation exoskeleton coupled with a musculoskeletal
 622 model via deep reinforcement learning. *Journal of neuroengineering and rehabilitation*, 20(1):34,
 623 2023.
 624

625 Anuj Mahajan, Mikayel Samvelyan, Lei Mao, Viktor Makoviychuk, Animesh Garg, Jean Kossaifi,
 626 Shimon Whiteson, Yuke Zhu, and Animashree Anandkumar. Tesseract: Tensorised actors for
 627 multi-agent reinforcement learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of*
 628 *the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine*
 629 *Learning Research*, pp. 7301–7312. PMLR, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/mahajan21a.html>.
 630

631 Thomas M Moerland, Joost Broekens, Aske Plaat, Catholijn M Jonker, et al. Model-based re-
 632 inforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 16(1):1–118,
 633 2023.
 634

635 Tung D Nguyen, Rui Shu, Tuan Pham, Hung Bui, and Stefano Ermon. Temporal predictive coding
 636 for model-based planning in latent space. In *International Conference on Machine Learning*, pp.
 637 8130–8139. PMLR, 2021.
 638

639 Ian Osband and Benjamin Van Roy. Near-optimal Reinforcement Learning in Factored MDPs. *arXiv*
 640 *e-prints*, art. arXiv:1403.3741, March 2014. doi: 10.48550/arXiv.1403.3741.
 641

642 Shubham Pateria, Budhitama Subagdja, Ah-hwee Tan, and Chai Quek. Hierarchical reinforcement
 643 learning: A comprehensive survey. *ACM Computing Surveys (CSUR)*, 54(5):1–35, 2021.
 644

645 Bei Peng, Tabish Rashid, Christian Schroeder de Witt, Pierre-Alexandre Kamienny, Philip Torr,
 646 Wendelin Böhmer, and Shimon Whiteson. Facmac: Factored multi-agent centralised policy
 647 gradients. *Advances in Neural Information Processing Systems*, 34:12208–12221, 2021.
 648

649 Alexandre Piché, Valentin Thomas, Cyril Ibrahim, Yoshua Bengio, and Chris Pal. Probabilis-
 650 tic planning with sequential monte carlo methods. In *International Conference on Learning*
 651 *Representations*, 2018.
 652

653 Thomas Pierrot, Valentin Macé, Jean-Baptiste Sevestre, Louis Monier, Alexandre Laterre, Nicolas
 654 Perrin, Karim Beguir, and Olivier Sigaud. Factored action spaces in deep reinforcement learning.
 655 2021.

648 Yuta Saito, Jihan Yao, and Thorsten Joachims. Potec: Off-policy learning for large action spaces via
 649 two-stage policy decomposition. [arXiv preprint arXiv:2402.06151](https://arxiv.org/abs/2402.06151), 2024.
 650

651 Pierre Schumacher, Daniel Häufle, Dieter Büchler, Syn Schmitt, and Georg Martius. Dep-rl: Embod-
 652 ied exploration for reinforcement learning in overactuated and musculoskeletal systems. [arXiv
 653 preprint arXiv:2206.00484](https://arxiv.org/abs/2206.00484), 2022.

654 Younggyo Seo, Carmelo Sferrazza, Haoran Geng, Michal Nauman, Zhao-Heng Yin, and Pieter
 655 Abbeel. Fasttd3: Simple, fast, and capable reinforcement learning for humanoid control. [arXiv
 656 preprint arXiv:2505.22642](https://arxiv.org/abs/2505.22642), 2025.

657 Carmelo Sferrazza, Dun-Ming Huang, Xingyu Lin, Youngwoon Lee, and Pieter Abbeel. Humanoid-
 658 bench: Simulated humanoid benchmark for whole-body locomotion and manipulation. [arXiv
 659 preprint arXiv:2403.10506](https://arxiv.org/abs/2403.10506), 2024.

660

661 Yunhao Tang and Shipra Agrawal. Discretizing continuous action space for on-policy optimization.
 662 In [Proceedings of the aaai conference on artificial intelligence](https://aaai.org/ocs/index.php/AAAI/AAAI20/paper/1683), volume 34, pp. 5981–5988, 2020.

663

664 Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden,
 665 Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. [arXiv preprint
 666 arXiv:1801.00690](https://arxiv.org/abs/1801.00690), 2018.

667 Arash Tavakoli, Fabio Pardo, and Petar Kormushev. Action branching architectures for deep rein-
 668 force learning. 2017.

669

670 Arash Tavakoli, Fabio Pardo, and Petar Kormushev. Action branching architectures for deep rein-
 671 force learning. In [Proceedings of the aaai conference on artificial intelligence](https://aaai.org/ocs/index.php/AAAI/AAAI20/paper/1683), volume 32,
 672 2018.

673 Tom Van de Wiele, David Warde-Farley, Andriy Mnih, and Volodymyr Mnih. Q-learning in enormous
 674 action spaces via amortized approximate maximization. [arXiv preprint arXiv:2001.08116](https://arxiv.org/abs/2001.08116), 2020.

675

676 Alexander Sasha Vezhnevets, Simon Osindero, Tom Schaul, Nicolas Heess, Max Jaderberg, David
 677 Silver, and Koray Kavukcuoglu. Feudal networks for hierarchical reinforcement learning. In [International conference on machine learning](https://proceedings.mlr.press/v70/vezhnevets17a.html), pp. 3540–3549. PMLR, 2017.

678

679 Zhihai Wang, Jie Wang, Qi Zhou, Bin Li, and Houqiang Li. Sample-efficient reinforcement learning
 680 via conservative model-based actor-critic. In [Proceedings of the AAAI Conference on Artificial
 681 Intelligence](https://aaai.org/ocs/index.php/AAAI/AAAI22/paper/1683), volume 36, pp. 8612–8620, 2022.

682

683 Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Mastering visual continuous control:
 684 Improved data-augmented reinforcement learning. [arXiv preprint arXiv:2107.09645](https://arxiv.org/abs/2107.09645), 2021a.

685

686 Denis Yarats, Amy Zhang, Ilya Kostrikov, Brandon Amos, Joelle Pineau, and Rob Fergus. Improving
 687 sample efficiency in model-free reinforcement learning from images. In [AAAI](https://aaai.org/ocs/index.php/AAAI/AAAI21/paper/1683), pp. 10674–10681,
 2021b.

688

689 Wenbo Zhang and Hengrui Cai. On high-dimensional action selection for deep reinforcement
 690 learning.

691

692 Mingde Zhao, Zhen Liu, Sitao Luan, Shuyuan Zhang, Doina Precup, and Yoshua Bengio. A
 693 consciousness-inspired planning agent for model-based reinforcement learning. [Advances in
 694 neural information processing systems](https://proceedings.neurips.cc/paper/2021/file/1569-1581.pdf), 34:1569–1581, 2021.

695

696 Dianyu Zhong, Yiqin Yang, and Qianchuan Zhao. No prior mask: Eliminate redundant action for
 697 deep reinforcement learning. In [Proceedings of the AAAI Conference on Artificial Intelligence](https://aaai.org/ocs/index.php/AAAI/AAAI24/paper/1683),
 698 volume 38, pp. 17078–17086, 2024.

699

700

701

APPENDIX

704 The technical appendix includes:

- 705 • Section A: A detailed description of the world model architecture and its learning objectives.
- 706 • Section B: An overview of the action space sizes for all visual control tasks used in our experiments.
- 707 • Section C: Additional quantitative and qualitative results, including:
 - 708 – Section C.1: Visualization of learned policies.
 - 709 – Section C.2: Comparison of the planning efficiency per environment interaction step with
 - 710 TD-MPC2, which is based on MPPI planning.
 - 711 – Section C.3: Visualizations of the learned action-dimensional preferences, demonstrating that
 - 712 the high-level policy generates focused and interpretable control-unit prompts.
 - 713 – Section C.4: Additional results of FOCUS on low-dimensional tasks, highlighting its versatility
 - 714 across diverse visual control settings.
 - 715 – Section C.5: Detailed experimental settings for `HumanoidBench`.
 - 716 – Section C.6: Hyperparameter sensitivity analyses on `DogWalk`.
 - 717 – Section C.7: Ablation studies on alternative high-level policy network designs that enforce
 - 718 explicit dependencies across action dimensions, rather than sampling them independently.
 - 719 – Section C.8: Full action-dimension selection frequencies, illustrating successful reduction of
 - 720 the effective action space.
 - 721 – Section C.9: Full empirical results on `HumanoidBench` with extensive 6 tasks, further demon-
 - 722 strating the efficiency of FOCUS’s hierarchical diversity-driven policy learning in high-
 - 723 dimensional action space.
- 724 • Section D: The description of LLM usage in this work.

A WORLD MODEL DETAILS

729 The DreamerV3 world model is implemented as a recurrent state-space model (RSSM), which
 730 maintains a deterministic hidden state h_t and a stochastic latent variable z_t to encode observations
 731 and predict future states. In our main text, we denote the concatenated feature $s_t = [h_t, z_t]$ as the
 732 overall state representation. The world model can be formulated as follows:

$$\begin{aligned}
 \text{Sequence model: } & h_t = f_\phi(h_{t-1}, z_{t-1}, a_{t-1}), \\
 \text{Encoder: } & z_t \sim q_\phi(z_t | h_t, x_t), \\
 \text{Dynamics predictor: } & \hat{z}_t \sim p_\phi(\hat{z}_t | h_t), \\
 \text{Reward predictor: } & \hat{r}_t = r_\phi(h_t, z_t), \\
 \text{Continue predictor: } & \hat{c}_t = c_\phi(h_t, z_t), \\
 \text{Decoder: } & \hat{x}_t = g_\phi(h_t, z_t),
 \end{aligned} \tag{7}$$

741 where f_ϕ is the recurrent transition function modeled by a GRU. q_ϕ and p_ϕ are implemented with
 742 MLPs. We use another MLP with separate heads r_ϕ and c_ϕ to predict the reward r_t and the episode
 743 continuation flag c_t , respectively. We use a CNN decoder to reconstruct the observation x_t .

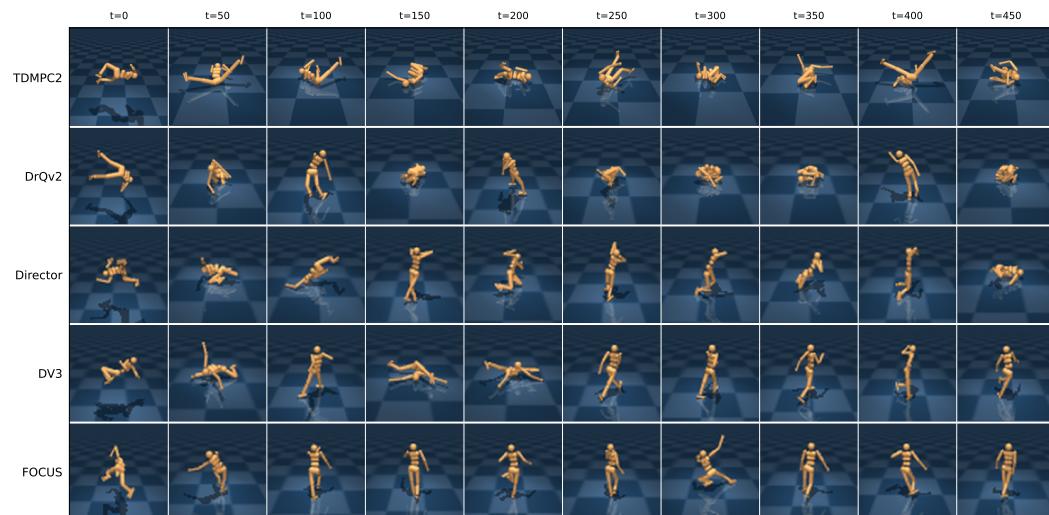
744 The world model is trained by minimizing a variational objective that combines reconstruction losses
 745 with KL regularization. In particular, the total loss is written as:

$$\begin{aligned}
 \mathcal{L}(\phi) &= \mathbb{E}_{\tau \sim \mathcal{B}} \left[\sum_{t=1}^T \left(\mathcal{L}_t^{\text{pred}} + \mathcal{L}_t^{\text{dyn}} + \mathcal{L}_t^{\text{rep}} \right) \right], \\
 \mathcal{L}_t^{\text{pred}} &= -\ln p_\phi(x_t | h_t, z_t) - \ln p_\phi(r_t | h_t, z_t) - \ln p_\phi(c_t | h_t, z_t), \\
 \mathcal{L}_t^{\text{dyn}} &= \text{KL}(\text{sg}(q_\phi(z_t | h_t, x_t)) \| p_\phi(\hat{z}_t | h_t)), \\
 \mathcal{L}_t^{\text{rep}} &= \text{KL}(p_\phi(\hat{z}_t | h_t) \| \text{sg}(q_\phi(z_t | h_t, x_t))).
 \end{aligned} \tag{8}$$

754 where $\mathcal{L}_t^{\text{pred}}$ is a self-supervised loss that reconstructs x_t and predicts the reward r_t and terminal
 755 signal c_t ; $\mathcal{L}_t^{\text{dyn}}$ is a KL divergence aligning the prior $p_\phi(\hat{z}_t | h_t)$ to the posterior $q_\phi(z_t | h_t, x_t)$; and
 $\mathcal{L}_t^{\text{rep}}$ is a reverse KL that regularizes the latent representations.

756 Table 1: Overview of action space dimensions for all continuous control tasks in our experiments.
757

758	Environment	Tasks	Action Space Size (\mathbb{R}^d)
759	High-dimensional DMC	Humanoid Stand	21
760		Humanoid Walk	21
761		Dog Walk	38
762		Dog Trot	38
763	MyoSuite	Key Turn Hard	39
764		Pen Twirl Hard	39
765		Pose Hard	39
766	Low-dimensional DMC	Walker Walk	6
767		Cheetah Run	6
768		Quadruped Walk	12
769			

770 Figure 9: Demonstration of the learned policy on HumanoidWalk.
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786787

B SUMMARY OF EXPERIMENTAL ENVIRONMENTS

788
789 Table 1 summarizes the environmental setups in the DeepMind Control and Myosuite benchmarks.
790 While FOCUS is designed to tackle high-dimensional problems with large action spaces, its results
791 on low-dimensional tasks highlight broad applicability.
792793

C ADDITIONAL QUANTITATIVE AND QUALITATIVE RESULTS

794

C.1 VISUAL EVALUATION OF LEARNT POLICIES

795
796 We evaluate trained policies of different models on the DMC and Myosuite tasks and visualize episode
797 frames for comparison. The full results on HumanoidWalk, DogTrot, and KeyTurnHard are
798 presented in Figure 9, Figure 10 and Figure 11 respectively.
799800

C.2 PLANNING EFFICIENCY COMPARISON WITH TD-MPC2

801 We compare the planning efficiency of FOCUS and TD-MPC2 in Figure 12, measuring the time
802 consumed from receiving an observation to producing an action. The experiments were conducted
803 on a system equipped with an NVIDIA RTX 4090 GPU and a dual-socket Intel Xeon Gold 6240C
804 CPU at 2.60GHz. For a fair comparison, both compared models are implemented in PyTorch 2.4.1,
805 whereas the proposed FOCUS is implemented in JAX for the main text and all other experiments.
806



Figure 10: Demonstration of the learned policy on DogTrot.

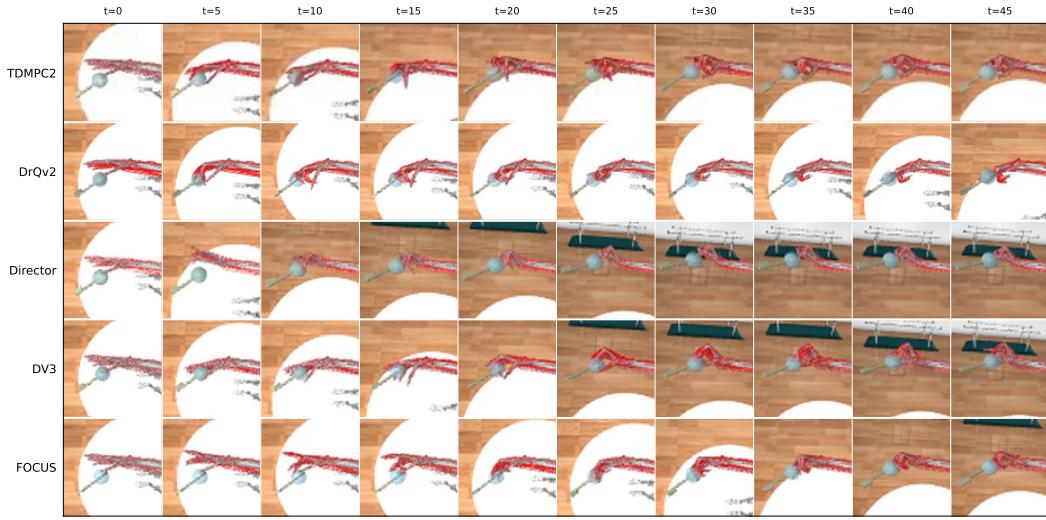


Figure 11: Demonstration of the learned policy on KeyTurnHard.

Although both methods rely on online planning, TD-MPC2 uses the *Model Predictive Path Integral* (MPPI) algorithm which is generally more computationally intensive than the proposed Sequential Monte Carlo (SMC) planning employed in FOCUS. As shown, FOCUS achieves significantly lower latency than TD-MPC2, with a notably smaller increase in latency across varying planning horizons. This also highlights the scalability of FOCUS for long-horizon tasks.

C.3 VISUALIZATION OF LEARNED HIGH-LEVEL PREFERENCES

We visualize the preference prompts sampled by the learned high-level policy in Figure 13, and present the full visualization video in supplementary materials. The results show that our hierarchical policy can selectively prioritize action dimensions corresponding to key joints, such as the legs and arms in the humanoid tasks.

For instance, in the second row in Figure 13, the policy focuses on the *left knee* and the *x-actuator at the right ankle* to support the stand-up motion. A closer look at joint control axes reveals that the selected x-actuator at the right ankle is oriented vertically to the ground, making it particularly effective for initiating a standing-up behavior. In the third row, upon detecting that the right lower leg requires adjustment, the policy prioritizes control of the right knee joint to enable stable locomotion.

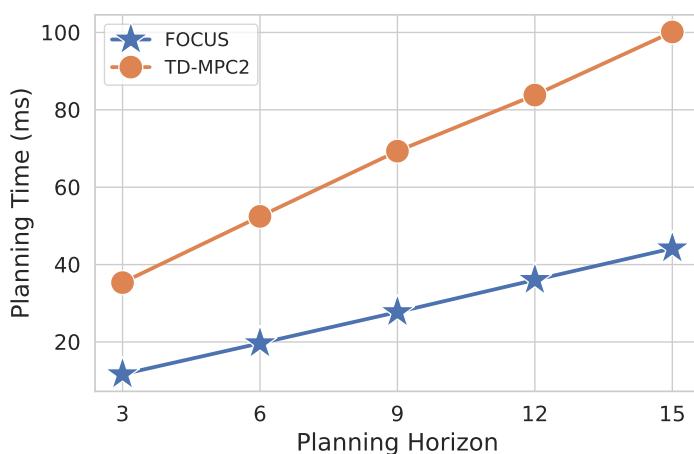


Figure 12: **Comparison of planning efficiency.** We measure the time per interaction step across different planning algorithms.

Table 2: **Episode return on DMC with low-dimensional action spaces.** The comparable performance to DreamerV3 highlights the applicability of FOCUS across diverse continuous visual control setups.

Method	Walker Walk	Cheetah Run	Quadruped Walk	Average
FOCUS	973.3 \pm 13.5	910.1 \pm 61.44	953.0 \pm 8.33	945.5
DreamerV3	966.4 \pm 21.19	901.9 \pm 26.2	943.4 \pm 29.1	937.2
TD-MPC2	332.9 \pm 31.9	202.7 \pm 2.3	162.2 \pm 49.7	232.6

C.4 APPLICABILITY TO TASKS WITH LOW-DIMENSIONAL ACTION SPACES

A key concern regarding the proposed FOCUS is whether the high-level preference policy might hinder sufficient exploration in environments with small action spaces. To examine this, we evaluate FOCUS on four tasks from the DeepMind Control Suite characterized by low-dimensional action spaces: Walker Walk, Cheetah Run, and Quadruped Walk, while additional task details can be found in Table 1. As shown in Table 2, FOCUS achieves performance comparable to DreamerV3 across these low-dimensional control tasks, demonstrating its broad applicability.

C.5 EXPERIMENTAL DETAILS ON HUMANOIDBENCH

Setups. The idea of diversity-driven hierarchical policy learning also extends to complex humanoid control tasks that use proprioceptive robot states as input. To validate this insight, we conduct additional experiments on two tasks from HumanoidBench (Sferrazza et al., 2024), namely H1hand-push and H1hand-reach.

Model implementations. We adopt the training framework of FastTD3 (Seo et al., 2025), which has been shown to efficiently solve HumanoidBench tasks through parallel simulation, large-batch updates, a distributional critic, and carefully tuned hyperparameters. We modify FastTD3’s policy architecture to incorporate our hierarchical policy model and diversity-driven objectives, resulting in a variant that we refer to as **FastTD3-FOCUS**. Concretely, we implement the high-level preference policy as a 3-layer MLP that outputs Bernoulli logits, while keeping FastTD3’s low-level policy architecture unchanged except for the ability to condition on preference prompts. During training, we compute both $\mathcal{L}_t^{\text{div}}$ and $\mathcal{L}_t^{\text{est}}$ in addition to FastTD3’s original policy loss. Since FastTD3 is a model-free algorithm, FastTD3-FOCUS follows the same setting and does not include world model learning or SMC planning.

More explanations on the results. Empirical results are presented in Figure 4 and Figure 16, comparing FastTD3 and FastTD3-FOCUS. This comparison helps illustrate the effect of our diversity-driven hierarchical policy learning objective, especially in tasks with relatively low-dimensional proprioceptive observations. While many HumanoidBench tasks (e.g., H1hand-walk,

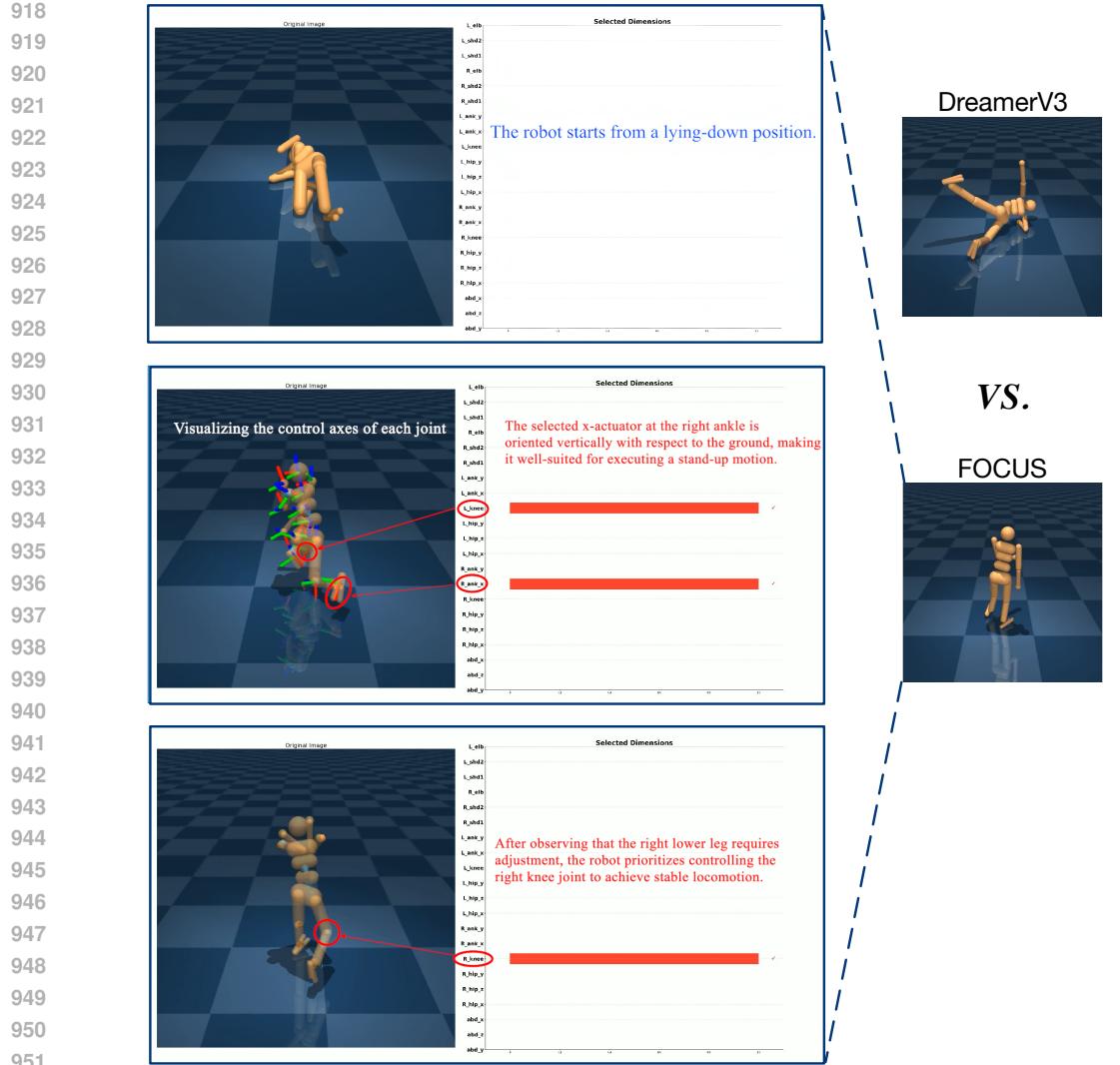


Figure 13: **Visualization of learned action-dimensional preferences.** Red bars indicate the action dimensions selected by the high-level policy. Please refer to our supplementary video for details.

H1hand-stand) become easy when trained using FastTD3’s framework, several tasks remain challenging due to their high-dimensional action spaces. On both H1hand-push and H1hand-reach, FastTD3-FOCUS exhibits improved sample efficiency over the vanilla FastTD3 baseline, which uses a simple MLP policy. Notably, since the FastTD3-FOCUS results do not incorporate high-level SMC planning, the improvements provide clean evidence of the effectiveness of our diversity-driven learning objective on its own.

C.6 HYPERPARAMETER ANALYSES ON DOGWALK

We further perform a hyperparameter sensitivity analysis on DogWalk, complementing the experiments on HumanoidWalk, with results shown in Figure 14(Left). As noted earlier, increasing N_p does not necessarily improve performance, since the exclusive computation of the policy log-likelihood constrains the advantage of sampling more preference candidates.

C.7 ABLATIONS ON CROSS-UNIT INTERDEPENDENCIES

We conduct ablation experiments in which dependencies across action dimensions are explicitly enforced, rather than sampled independently. Specifically, we apply a Softmax to the logits output by

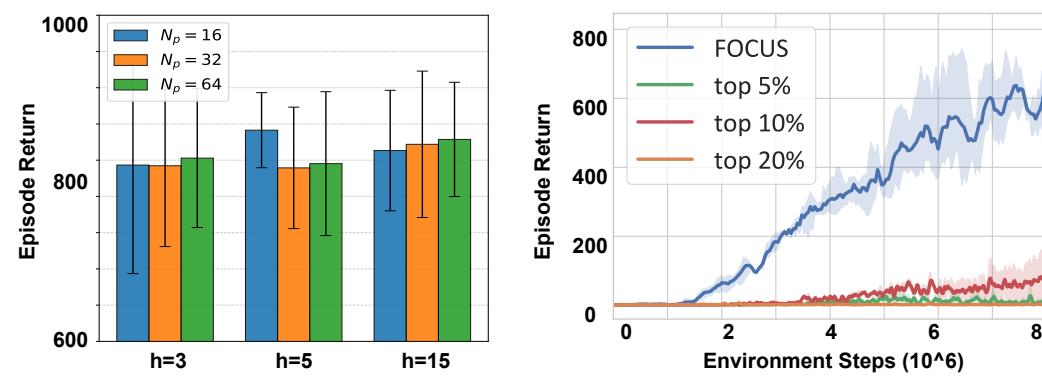


Figure 14: **Left:** Ablation study results on DogWalk. We evaluate the Effect of planning horizon h and candidate trajectories N_p in our planner. **Right:** Comparison of approaches that explicitly model inter-unit correlations. Test curves are averaged over three random seeds on HumanoidWalk.

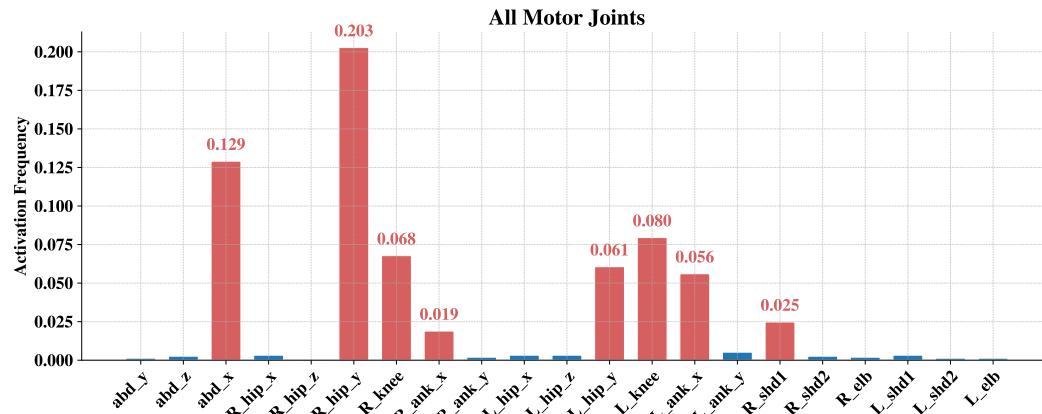


Figure 15: **Action-dimension selection frequencies.** FOCUS effectively leverages structured sparsity to guide focused exploration, supporting our objective of reducing exploration overhead in high-dimensional control tasks.

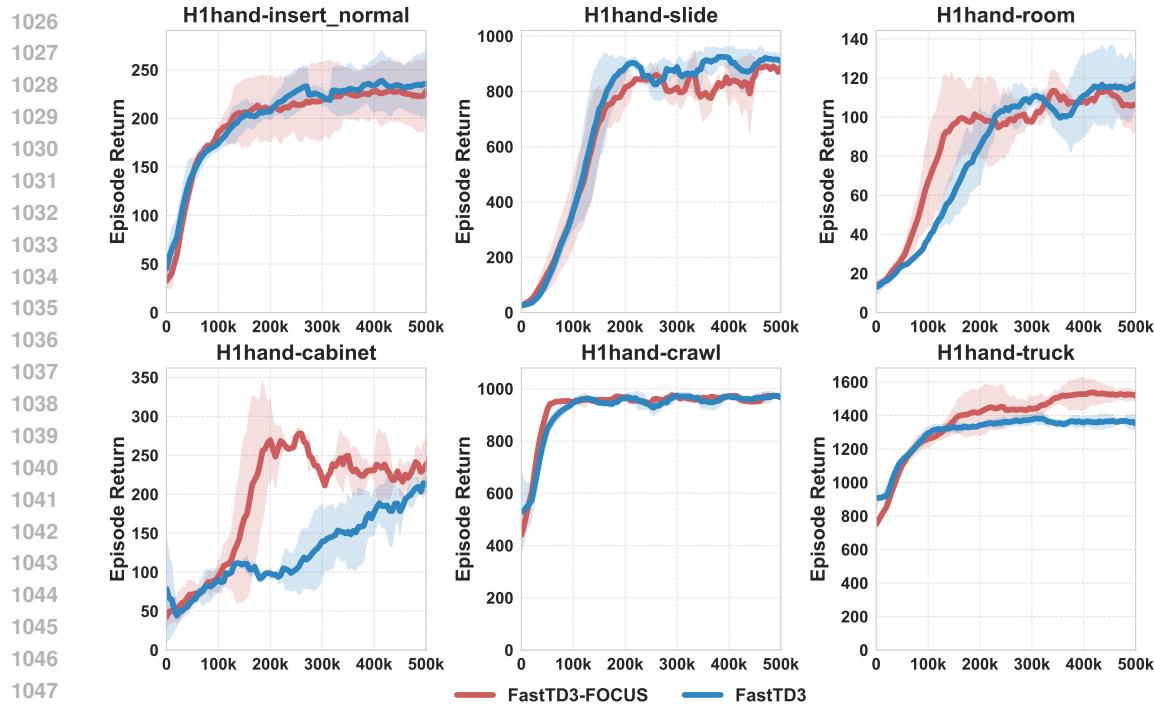
the high-level preference policy, followed by a Gumbel top- $k\%$ selection to determine the preference prompts. We test Top-5%, Top-10%, and Top-20% settings. The results, shown in Figure 14(Right), indicate that despite tuning the hyperparameter k , this design consistently fails to learn meaningful behaviors on the HumanoidWalk task. In contrast, our approach enables the low-level policy to implicitly capture correlations across control units, yielding more effective policy learning.

C.8 FULL ACTION-DIMENSION SELECTION FREQUENCIES

We present the full action-dimension selection frequency in Figure 15, complementing Figure 7. Actuators beyond the top 8 most frequently selected ones are rarely activated, indicating that the high-level preference policy effectively discovers structural sparsity and reduces the action space.

C.9 FULL RESULTS ON HUMANOIDBENCH

We present here the full empirical results on 6 additional tasks in HumanoidBench. The results in Figure 16 further demonstrate that FOCUS achieves significantly higher sample efficiency (e.g., cabinet, room) and the capability to avoid local optima (e.g., truck) in high-dimensional action spaces. For tasks like crawl and slide, our method maintains competitive performance comparable to the baseline.

Figure 16: **Episodic returns of 6 tasks on HumanoidBench.**

D USE OF LLMs

Large language models (LLMs) were used exclusively for language editing, including grammar correction and sentence polishing. They were not used for developing ideas, designing methods, conducting experiments, or interpreting results. Specifically, we provided prompts such as "Help polish this sentence to improve clarity and fluency in an academic writing style.", "Rewrite this sentence in a different expression", or "What is the alternative word for xxx". The output was then evaluated and further edited to stay close to this paper's writing style.