2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

VB-Com: Learning Vision-Blind Composite Humanoid Locomotion Against Deficient Perception

Junli Ren [1,2] Tao Huang[1] Huayi Wang[1] Zirui Wang[1] Qingwei Ben[1] Junfeng Long[1]
Yanchao Yang[2] Jiangmiao Pang[1,†] Ping Luo[2,†]
[1]Shanghai AI Laboratory [2]The University of Hong Kong [†]Equal Advising

Fig. 1: Overview. VB-Com enables humanoid robots (move direction in orange arrorw) to traverse dynamic disturbances (move direction in blue arrorw), including (a) suddenly appearing obstacles/hurdles, (b) deformable gaps, and (c) sensor occlusions. We demonstrate the effectiveness of VB-Com on vairous humanoid robots of Unitree G1 and H1.

***Abstract*— The performance of legged locomotion is highly dependent on the accuracy and completeness of state observations. While perceptive locomotion enables robots to plan motions proactively and adapt to unstructured terrains, real-world perception is often degraded by sensor noise, dynamic disturbances, or training outliers. These inaccuracies can lead to locomotion failures, particularly for humanoid robots, which are especially vulnerable to misguidance from imperfect perception. However, exhaustively simulating all potential perceptual deficiencies (e.g., dynamic or deformable terrains) during training remains impractical due to inherent simulation limitations. To address this fundamental challenge, we propose VB-Com - a novel framework that dynamically combines a vision-based policy (utilizing exteroceptive sensing) with a proprioception-only blind policy through intelligent composition. When inaccurate perception begins to destabilize locomotion, VB-Com is able to identify these degradation scenarios and immediately turns to "blind" actions and recovers from the potential failure. Our approach mitigates the risks posed by deficient perception that has not be addressed by existing research on perceptive locomotion. Experimental results demonstrate that VB-Com robustly enables humanoid robots to traverse challenging terrains and obstacles–under perceptual deficiencies and dynamic disturbances. Details demonstrations can be found in our project Website: vbcom.github.io.**

I. Introduction

While legged locomotion control has been well-addressed through reinforcement learning with effective data collection [1, 2] and well-crafted reward guidance [3–9], the performance of such policies remains highly dependent on the accuracy and comprehensiveness of state observations [9–12]. The state space can be roughly categorized into three types: 1) Accessible states, which are reliable and obtainable on real robots, such as joint encoders and IMU; 2) Privileged states [13, 14], which are unavailable on real

robots, including velocity and static hardware parameters; and 3) External states [7, 8, 15–17], which are observable but inherently noisy and occasionally unreliable. Previous "blind" locomotion policies [3, 12–14, 18] require robots to physically interact with unstructured terrains before responding, forcing a trade-off between sacrificing speed to ensure safety or acting quickly but failing in scenarios that demand rapid responses. On the other hand, Despite the impressive results imposed by perceptive locomotion [19–21] that enables the robot to anticipate incoming terrains and plan motions in advance, these methods are heavily dependent on maintaining consistency between perceived external states and those appeared during simulation [22]. When mismatches occur, the robot may exhibit abnormal or dangerous behaviors.

In practice, it is impossible to provide the robot with all potentially encountered external states within the simulator [23]. Current contact models in simulators are limited to rigid-body interactions and it is computationally expensive to incorporate dynamic terrains and obstacles during training [24]. Although previous research has highlighted the combination of perception and proprioception to achieve robust locomotion performance [15, 25, 26] against perception inaccuracy, these studies have primarily focused on quadruped robots and low-risk scenarios, where a delayed response to the environment does not typically lead to termination.

Despite the impressive results of recent research achieving humanoid motions through tele-operation and imitation learning [27–32], the bipedal lower-limb structure of humanoid robots presents unique challenges in locomotion control compared to quadrupeds [5, 33]. The shifting of the gravity center in humanoid robots makes them more prone to unrecoverable falls. As a result, humanoid robots are more vulnerable to unexpected disturbances. Consequently, current perceptive humanoid locomotion studies are limited to static terrains and confined environments, with performance heavily reliant on the quality of the perception module [8, 16].

In this work, we propose **VB-Com** (Vision-Blind Composite Humanoid Control), a hierarchical locomotion system capable of handling deficient perception out of training distribution and dynamic disturbances. VB-Com consists of two sub-modules: 1) Locomotion policies: A perceptive and a non-perceptive humanoid locomotion policy that can traverse gaps, hurdles and avoid obstacles. 2) Return Estimators: Two hardware-deployable return estimators that predicts future returns obtained by each locomotion policy conditioned on proprioceptive states observations. VB-Com dynamically determines whether to trust perception or disregard them to prevent locomotion failures caused by misleading observations. As shown in Fig. 1, we have provided extensive hardware demonstrations on VB-Com autonomously recovers the robot from deficient perception-induced instability.

Table I outlines VB-Com's advantage among prior works : (1) **robust** locomotion against **perception deficiency**, generalizing **beyond the noise distribution within training**. (2) resilience to **unperceptive disturbances** such as rushing pedestrians or deformable terrains. (3) We introduce obstacle avoidance at the locomotion level, a feature not addressed by previous humanoid locomotion works.

TABLE I: VB-Com v.s. Pervious Works

Method	Robust Perception	Generalizable beyond Training Noise	Unperceptive Disturbances	One-Stage Training	Avoid Obstacles	Traverse Gaps	Traverse Hurdles
Robust Perceptive [15]	✓	✗	✗	✗	✗	✗	✓
Humanoid Parkour [16]	✗	✗	✗	✗	✗	✓	✓
PIM [8]	✗	✗	✗	✓	✗	✓	✗
VB-Com (ours)	✓	✓	✓	✓	✓	✓	✓

II. Related Work

A. Robust Perceptive Legged Locomotion

Generally, lidar-based elevation maps [8, 22, 34] and depth images [19, 20, 35] are widely implemented in perceptive locomotion. However, depth images are significantly affected by lighting conditions and limited field of view, while lidar-based elevation maps are restricted to static environments. Later studies have focused on integrating proprioceptive and exteroceptive observations to achieve robust locomotion or navigation against deficient perception [15, 25, 26, 36, 37]. These approaches either use a belief encoder combining exteroceptive and historical proprioceptive data (limited by perception noise encountered in training) or handle perception errors at the path planning level (too slow for sudden disturbances). We address these challenges through dynamic policy composition: when deficient perception disrupts locomotion, VB-Com activates a proprioceptive-only "blind policy." Both policies share the same state/action spaces, enabling fast, stable recovery. Crucially, by detecting perception failures through proprioception rather than visual cues, VB-Com generalizes to unseen scenarios beyond its training distribution.

B. Hierarchical Reinforcement Learning

Hierarchical reinforcement learning has been extensively explored in the literature, with the composition of low-level skills emerging as a popular approach for addressing long-horizon or complex tasks [38–40]. Among these works, value functions play a crucial role in policy composition [41–43], particularly in capturing the affordances of each sub-task. VB-Com draws inspiration from these approaches by training two return estimators, each representing the capabilities of the vision and blind policies, respectively.

In addition, several works in legged locomotion have explored hierarchical structures, such as employing DAgger to distill a set of locomotion skills [44]. Recent research [45] also proposed a switching mechanism to achieve high-speed locomotion while avoiding obstacles. However, these frameworks rely heavily on vision observations, making them intolerant to perception outliers. In contrast, VB-Com addresses the novel challenge of maintaining stable locomotion despite deficient perception, with a specific focus on humanoid robots and high-dynamic tasks.

III. Preliminaries

A. Probelm Formulation

Reinforcement Learning based locomotion control is commonly modeled as a Partially Observable Markov Decision Process (POMDP), characterized by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{R})$. The control policy $\pi(a|o)$, typically represented by a neural

network, maps observations $o \in \mathcal{O}$ to actions $a \in \mathcal{A}$. Given the reward functions $r \in \mathcal{R}$ and a discount factor γ, the policy is trained to maximize the expected cumulative return:

$$J(\pi) = \mathbb{E}_{a_t \sim \pi(o_t)}[\sum_{t}^{\infty} \gamma^t r(s_t, a_t)], \tag{1}$$

in this work, we address a more challenging POMDP task where the partial observations $\mathcal{O}$ include a potentially unreliable component o_v, which can fail under certain conditions, leading to significant penalties or termination. However, completely discarding o_v would substantially limit the performance upper bound. The ideal solution is to enable the policy to recognize when o_v becomes unreliable and switch to relying solely on the reliable proprioceptive observations o_p. We propose a composite solution to address this challenge in this work.

B. Q-informed Policies Composition

Given a set of policies $\Pi = \{\pi_1, \pi_2, \ldots, \pi_n\}$ that share the same states, actions, and rewards $(\mathcal{S}, \mathcal{A}, \mathcal{R})$, a composite policy $\tilde{\pi}$ selects an action from the proposed action set $A = \{a_i \sim \pi_i(s)\}$ with a probability P_w that is related to their respective potential utilities. In the context of Markov Decision Process, it has been proved [43] that selecting actions based on the cumulative return at current states and candidate actions will achieve the best expected return for $\tilde{\pi}$. To this end, the Q-value based policies composition will compute the cumulative return at current for each low-level policy $\mathbf{Q} = \{Q_i(s, a_i) | a_i \in \mathcal{A}\}$ and construct a categorical distribution to select the final action:

$$P_w(i) = \frac{\exp(Q_i(s, a_i)/\alpha)}{\sum_j \exp(Q_j(s, a_j)/\alpha)}, \tag{2}$$

here α is the temperature. In the case of two sub-policies, such composition will assign a higher probability to the action with the higher Q-value at the current state.

IV. Method

A. System Overview

The proposed VB-Com framework (Fig. 2) comprises a perceptive locomotion policy π_v and a non-perceptive policy π_b. π_v incorporates visual observations to enable perceptive locomotion, π_b is trained within the same reward and action space but does not accept vision input. Both π_v and π_b are trained to operate stably on different terrains within the training distribution. During deployment, VB-Com primarily selects actions from the vision policy, leveraging its richer environmental observations and higher expected returns. However, when encountering perceptual outliers (e.g., environments that contradict vision-based observations), the system switches to the blind policy, relying on more reliable proprioceptive observations. This composition is enabled by two return estimators (π_v^e and π_b^e), trained jointly with the locomotion policies. At each timestep, the compositor evaluates estimated returns $\{\hat{G}_v^e \sim \pi_v^e, \hat{G}_b^e \sim \pi_b^e\}$ to select between candidate actions $\{a_v \sim \pi_v, a_b \sim \pi_b\}$.

B. Locomotion Policies

To demonstrate the quick responsiveness of VB-Com in handling deficient perception, we present the robot with challenge terrains including gaps, hurdles, and high walls (for obstacle avoidance). Both vision and blind policies are trained from scratch with same reward (Table II) and action space. The observation space includes proprioceptive states (joint positions/velocities, gravity, base orientation) with robot-centric heightmaps (for π_v only), while the critic involves privileged information (precise velocity, larger heightmaps). Based on the training, the vision policy executes vision-guided maneuvers including gap jumping, hurdle traversal, and collision-free obstacle avoidance, while the blind policy demonstrates comparable terrain traversal capability, when proprioception detects physical contact or balance loss.

TABLE II: Rewards

Reward	Equation	Weight: H1	Weight: G1
Task Rewards			
Tracking Goal Velocity	$\min(v_c, \|\mathbf{v}_{xy}\|)/v_c$	5.0	2.0
Tracking Yaw	$\exp\{(\mathbf{p}-\mathbf{x})/\|\mathbf{p}-\mathbf{x}\|\}$	5.0	2.0
Collision	$\sum_{i \in C_f^e} \mathbf{1}\{\|\mathbf{f}_i\| > 0.1\}$	-15.0	-15.0
Regularization Rewards			
Linear velocity (z)	v_z^2	-1.0	-1.0
Angular velocity (xy)	$\|\boldsymbol{\omega}_{xy}\|_2^2$	-0.05	-0.05
Orientation	$\|\mathbf{g}_x\|_2^2 + \|\mathbf{g}_y\|_2^2$	-2.0	-2.0
Joint accelerations	$\|\ddot{\boldsymbol{\theta}}\|_2^2$	-2.5×10^{-7}	-2.5×10^{-7}
Joint velocity	$\|\dot{\boldsymbol{\theta}}\|_2^2$	-5.0×10^{-4}	-5.0×10^{-4}
Torques	$\|\frac{\boldsymbol{\tau}}{k_p}\|_2^2$	-1.0×10^{-5}	-1.0×10^{-5}
Action rate	$\|\mathbf{a}_t - \mathbf{a}_{t-1}\|_2^2$	-0.3	-0.3
Joint pos limits	$\text{RELU}(\boldsymbol{\theta} - \boldsymbol{\theta}^{max}) + \text{RELU}(\boldsymbol{\theta}^{min} - \boldsymbol{\theta})$	-2.0	-2.0
Joint vel limits	$\text{RELU}(\dot{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}^{max})$	-1.0	-1.0
Torque limits	$\text{RELU}(\hat{\boldsymbol{\tau}} - \hat{\boldsymbol{\tau}}^{max})$	-1.0	-1.0
Motion Style Rewards			
Base Height	$(h - h^{\text{target}})^2$	-0.0	-10.0
Feet Air Time	$\sum_{i \in \text{feet}}^2 (t_{\text{air},i} - 0.5) \cdot \mathbf{1}\{\text{first ground contact}\}$	4.0	1.0
Feet Stumble	$\sum_{i \in \text{feet}} \mathbf{1}\{\|\boldsymbol{f}_i^{xy}\| > 3\|\boldsymbol{f}_i^z\|\}$	-1.0	-1.0
Arm joint deviation	$\sum_{i \in \text{arm}} \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^{\text{default}}\|^2$	-0.5	-0.5
Hip joint deviation	$\sum_{i \in \text{hip}} \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^{\text{default}}\|^2$	-5.0	-0.5
Waist joint deviation	$\sum_{i \in \text{waist}} \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^{\text{default}}\|^2$	-5.0	-0.0
Feet distance	$(\|\mathbf{p}_{\text{left foot}} - \mathbf{p}_{\text{left foot}}\| - d_{\min})$	1.0	0.0
Feet lateral distance	$\left(\|\mathbf{p}^y_{\text{left foot}} - \mathbf{p}^y_{\text{right foot}}\| - d_{\min}\right)$	10.0	0.5
Knee lateral distance	$\left(\|\mathbf{p}^y_{\text{left knee}} - \mathbf{p}^y_{\text{right knee}}\| - d_{\min}\right)$	5.0	0.0
Feet ground parallel	$\sum_{i \in \text{feet}} \text{Var}(\boldsymbol{p}_i^z)$	-10.0	-0.02

C. Vision-Blind Composition

Given the vision policy π_v and the blind policy π_b, the composition can be viewed as a discrete policy $\tilde{\pi}$ with an action dimension of two, selecting between the candidate actions:

$$\tilde{\pi}(a|s) = [a_b \sim \pi_b, a_v \sim \pi_v]\mathbf{w}, \mathbf{w} \sim P_{\mathbf{w}}, \tag{3}$$

Building on the analysis of Q-informed policy composition, for each state s_t at each step, we have:

$$P_{\mathbf{w}}(i|s_t, a_v, a_b) \propto \exp(Q(s_t, a_i)), a_i \in \{a_v, a_b\}. \tag{4}$$

1) Policy Return Estimation: Given the current states s_t of the robot, we can estimate the expected cumulative return $G_{\pi_i}(s_t)$ for each policy to guide the composition process. To avoid abrupt changes in the action space caused by frequent switches, we introduce a switch period T, which acts as the control unit for each switch. The introduction of T also helps decouple the switching actions, approximately making them temporally independent of each other.

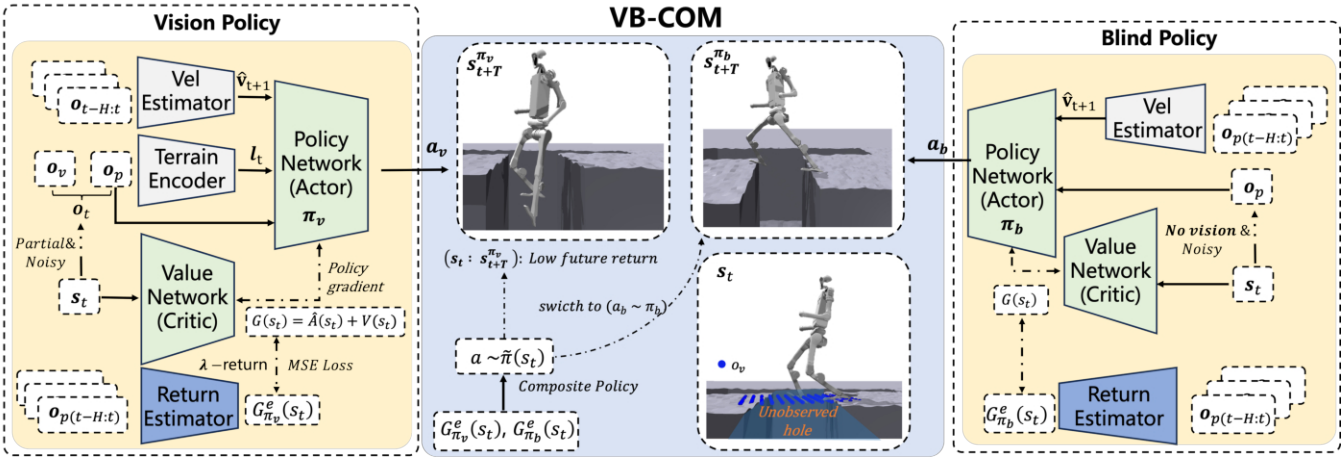

Fig. 2: Overview of our framework: In VB-Com, we develop two locomotion policies—one perceptive and one non-perceptive—through single-stage training. These sub-policies are integrated based on two return estimators, which predict future returns given the current state for each of the policy policy. This integration enables seamless policy switching, allowing the robot to effectively adapt to varying levels of perceptual deficiency and dynamic environments.

To this end, we expect the return estimator to be responsible for estimating a time sequence of expected returns over the duration T, such that:

$$L_{\pi_i} = \mathbb{E}_t[\hat{G}^e_{\pi_i}(s_t) - G_{\pi_i}(s_{t:t+T})], \tag{5}$$

To achieve the estimation with reduced bias and variance, we implement λ-return to weight the time-sequenced returns within one switch period as follows:

$$G_{\pi_i}(s_{t:t+T}) \approx G^{\lambda}_{\pi_i}(s_{t:t+T}) = (1-\lambda)\sum_{n=t}^{t+T}\lambda^{n-1}G_{\pi_i}(s_n), \tag{6}$$

which represents a weighted return if the robot chooses to switch to the low-level policy π_i given the state s_t. In addition, in order to mitigate the large variance between single-step rewards and prevents the policy from overfitting to recent batches, G_{π_i} is computed based on the update of value functions [46], where $G_{\pi_i}(s_t) = \hat{A}(s_t) + V(s_t)$, with $\hat{A}(s_t)$ being the advantage function and $V(s_t)$ the value function.

Since the return estimators need to be deployable on hardware and we aim to mitigate perception misleadings, we avoid using exteroceptive observations or privileged information as inputs. Instead, we use the historical proprioceptive observation sequence $o_{p_{t-H:t}}$ as the input to the return estimator π^e. In Section V-E, we investigate the superiority of the TD-based return estimator and different choices of the switching period within the system.

2) Policy Switch: Unlike previous works that construct a switch-based hierarchical framework to keep the robot within a safe domain and prevent potential collisions, VB-Com performs policy switching to recover the robot from getting stuck due to perceptive deficiencies.

Ideally, Eq. (4) provides the theoretical basis for choosing the action with the greater value estimation $\hat{G}^e_{\pi}$ at the current state. This aligns with the fact that π_v typically yields higher returns than π_b as long as the vision observations are consistent with those seen during training, since π_v has access to more comprehensive environmental observations.

During deployment, when the robot experiences a sudden environmental change that disrupts locomotion, both estimations $\hat{G}^e_{\pi_{v,b}}$ will decline. We observe that in these situations, it is difficult to maintain strict monotonicity such that $\hat{G}^e_{\pi_b} > \hat{G}^e_{\pi_v}$ due to the return approximation error introduced by π^e. Meanwhile, the blind policy demonstrates greater sensitivity to unstable motions, as the low-return samples are more frequently encountered even after the policy has been well-trained, compared to π_v (as illustrated in Fig. 3). To address this, we introduce a threshold G_{th} trigger that can also prompt the policy switch.

$$a \sim \tilde{\pi}(s_t) = \begin{cases} a_v, & \text{if } G^e_{\pi_v}(s_t) > G^e_{\pi_b}(s_t) > G_{th}, \\ a_b, & \text{otherwise,} \end{cases} \tag{7}$$

$$G_{th} = 1/5\sum_{t-5}^{t} G^e_{\pi_b}(s_i) - \alpha, \tag{8}$$

here α is a threshold hyperparameter. In practice, we replace $G^e_{\pi_v}(s_t)$ with a smoothed window (length 5) to avoid sudden abnormal estimations, which we have found to be effective in real robot deployments. Additionally, a switch will not be performed under conditions of high joint velocity to prevent potential dangers caused by the abrupt switching of policies when the robot is performing vigorous motion. Section V-E also discusses the effectiveness of different α within the system.

V. Evaluations

A. Implementation Settings

We deploy VB-Com on two humanoid platforms—Unitree G1 and H1—in both simulation and real-world settings. G1 is controlled via 20 joints (10 upper-body, 10 lower-body), and H1 via 19 joints (8 upper-body, 10 lower-body, and 1 torso). Both robots use whole-body control and share a common robot-centric elevation map (see Fig. 4) to provide external observations for the vision policy. Onboard lidars mounted on the head serve as the primary sensing modality. To improve the robustness of the base policy π_b under perceptual

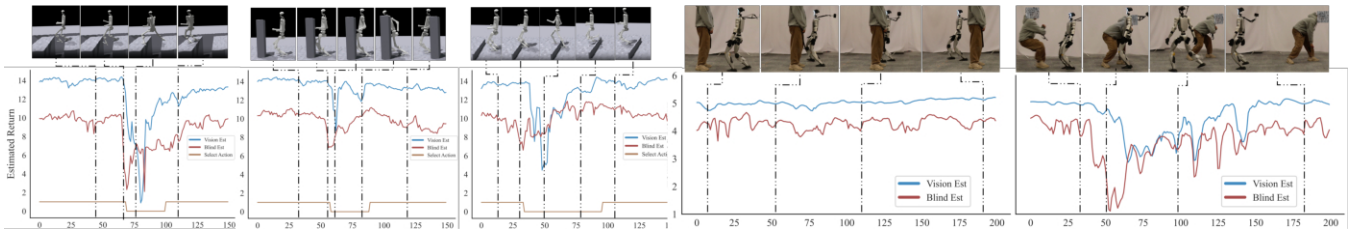

Fig. 3: Illustrations of the variation in estimated return and action phases(0 for a_b and 1 for a_v) across three concerned terrains.

TABLE III: VB-Com Evaluations on graded perceptual deficiency against baselines

Noise Level	Robust Perceptive [15]		Vision Policy		VB-Com	
	Goals Completed(%)	Collision Steps(%)	Goals Completed(%)	Collision Steps(%)	Goals Completed(%)	Collision Steps(%)
0%	80.33 ± 5.45	4.16 ± 1.45	73.57 ± 4.97	1.39 ± 0.53	84.05 ± 2.28	1.50 ± 0.14
30%	78.14 ± 1.62	3.25 ± 1.02	72.76 ± 2.29	2.52 ± 0.32	82.25 ± 6.6	2.09 ± 0.13
70%	56.62 ± 1.78	7.07 ± 0.98	55.38 ± 3.33	6.08 ± 0.82	82.48 ± 1.20	2.12 ± 0.11
100%	51.29 ± 9.20	4.66 ± 0.43	48.71 ± 5.60	6.92 ± 1.36	83.76 ± 1.35	2.57 ± 0.27

Proprioceptive Observations	**Dimension**
Dof Position	19(H1) / 20(G1)
Dof Velocity	19(H1) / 20(G1)
Projected Gravity	3
Base Angular Velocity	3
Action (Dof Position)	19 / 20
Locomotion Frequency	50hz
Hyperparameters	**Values**
Perception Sensor	Lidar(Mid360)
Perception Frequency	10hz
Heightmap Range Forward (m)	$[-0.35, 0.85]$
Heightmap Range Lateral (m)	$[-0.35, 0.35]$
Velocity Command Range (m/s)	$[0.0, 1.0]$
Yaw Command Range (rad/s)	$[-0.5, 0.5]$
Curriculum	**Ranges** (TL: Terrain Level)
Gap Width Curriculum Range (m)	$[0.1 + 0.5 * TL, 0.2 + 0.6 * TL]$
Hurdles Heights Curriculum Range (m)	$[0.1 + 0.1 * TL, 0.2 + 0.2 * TL]$
Obstacles Length Curriculum Range (m)	$[0.1 + 0.1 * TL, 0.2 + 0.2 * TL]$

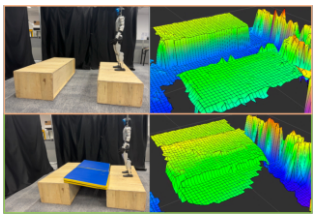

Fig. 4: (Left) Implementation Details. (Right) We demonstrate the hardware perception under ideal and deficient perception situations.

degradation, we introduce standard training-time noise (see Fig. 5), including 10% Gaussian noise and random perception delays of up to 0.5 seconds—resulting in a vision policy that is more resilient and deployable in practice.

B. Example Case

Fig. 3 demonstrates how VB-Com responds to perceptual deficiencies in both simulation and real-world settings. In simulation, when walking on flat terrains, the vision policy initially dominates as its estimated return $G^e_{\pi_v}$ remains higher than $G^e_{\pi_b}$. When deficient perception fails at responding to challenging terrain, $G^e_{\pi_v}$ drops sharply, triggering an immediate switch to the blind policy π_b. This fallback stabilizes the motion until stable motion recovers, at which point VB-Com smoothly transitions back to π_v.

A similar pattern emerges on real hardware. For static obstacles, the elevation mapping allows the vision policy to remain active, maintaining a high estimated return. In contrast, when a fast-moving person approaches the robot and the perception module fails to detect it in time, both returns decline, prompting a timely switch to π_b for reactive collision avoidance.

These results highlight VB-Com's responsiveness and robustness under degraded perception, both in simulation and the real world.

C. Evaluations on Graded Perceptual noises

To evaluate VB-Com's ability to handle perception deficiencies, we design experiments **simulating increasingly severe perceptual noise** across three challenging terrain types (the noises illustrated in Fig. 5), we conduct 10 trials per condition with 3 random seeds for statistical reliability. Each trial requires the robot to navigate 8 goal points paired with challenge terrains/obstacles. We compare VB-Com against standalone vision/blind policies and baselines from prior work: Robust Perceptive policy [15] that addresses perception deficiency by estimating external states from **graded perception noises** (Table III). We also include a Noisy Perceptive Policy baseline (trained with evaluation noises) to compare the performance limits of traditional noise augmentation versus VB-Com's proprioception-driven switching for novel perception failures (Section V-D).

The results reveal critical limitations of current approaches - while the Robust Perceptive policy slightly outperforms VB-Com's vision component on seen noises, its performance collapses by **40**% when encountering training-unseen perception failures (Fig. 5-(left)). VB-Com overcomes these limitations through its unique switching mechanism, maintaining **82**%+ success rates across all noise levels by leveraging proprioceptive cues to detect and recover from unexpected perception failures. The results demonstrate that unlike methods requiring prior exposure to specific failure modes, **VB-Com's proprioception-driven switching enables reliable operation even with novel perception deficiencies and dynamic disturbances**.

Fig. 5 provides a more intuitive illustration investigating the advantage of VB-Com over standalone blind policy and other perceptive baselines: **as perception becomes more comprehensive, VB-Com achieves both fewer collisions and better goal-reaching performance**. In contrast, the blind policy maintains a high goal-reaching rate but results in more collisions, while the vision policy performs better in avoiding collisions when the perception is accurate and comprehensive. **As the noise level increases, the performance of VB-Com begins to resemble that of the blind policy.** These results demonstrate the effectiveness of the composition system, which benefits from both sub-policies to achieve better

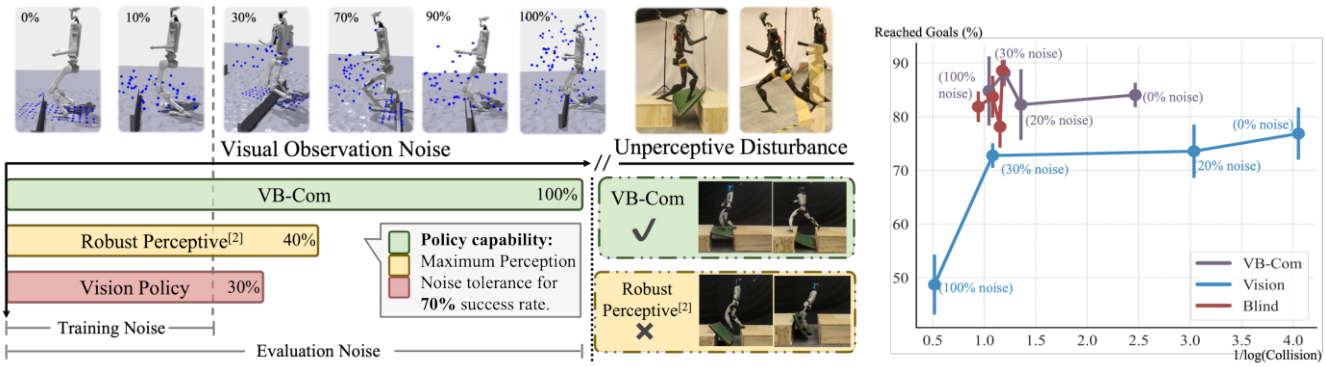

Fig. 5: Illustrations of the capability of VB-Com in continuous noisy space.

TABLE IV: Ablations with switch thresholds.

	w/o Thresholds $G^e_{\pi_v}(s_t) > G^e_{\pi_b}(s_t)$	**w/o Blind Return** $G^e_{\pi_v}(s_t) > G_{\text{th}}$	**Default** $\alpha = 1.0$	$\alpha = 2.0$	$\alpha = 0.5$	$\alpha = 0.1$
Goals Completed (%)	48.48 ± 1.28	76.95 ± 2.45	84.05 ± 2.28	77.10 ± 4.71	85.76 ± 2.88	84.43 ± 1.23
Collision Steps (%)	6.24 ± 0.41	4.06 ± 0.16	1.50 ± 0.14	2.63 ± 0.68	2.29 ± 0.17	2.10 ± 0.25

performance in terms of both goal-reaching and minimizing collisions.

D. Comparisons with Noisy Perceptive Training

To further demonstrate VB-Com's unique advantage in overcoming vision deficiencies through proprioceptive observations, we introduce a **Noisy Perceptive Policy baseline** trained with the **evaluation noises** in a curriculum matching the terrain difficulty. As shown in Fig. 6-(a,b), the experiments reveal VB-Com's superior generalization capability - proving that **simply incorporating noise in training cannot match VB-Com's ability to handle novel perception failures through its proprioception-driven switching mechanism.**

E. Return Estimator Evaluations

Given the critical role of the return estimator in VB-Com, we conduct a comprehensive analysis of its performance across different design choices.

VB-Com tolerates noisy return estimates through this threshold-based switching mechanism. We observe that under perceptually degraded conditions, return comparisons become less reliable due to elevated estimation errors (Fig. 8). To address this, we introduce a return threshold parameter G_{th}, which ensures robust and timely switching even when return estimates are noisy (Table IV). Table IV also shows that system performance is not sensitive to the choice of the hyperparameter α, supporting the robustness and generalizability of the proposed method.

The proposed TD-based return estimator within the vision policy convergent stably as it updates alongside the locomotion policy. Since we update the return estimator using temporal difference, we compare it with the Monte Carlo-based search return estimator that estimate the furtuen expected returns with the following regression loss directly: $\mathbb{E}_t[\hat{G}^e_{\pi_i}(s_t) - \sum_t^{t+T} \gamma^t r(s_t, a_t)]$. As shown in Fig. 6-(d), the MC-based estimator struggles to converge due to the accumulation of noise.

TABLE V: Ablations without repeating metric labels.

Method	Goals Completed(%)	Collisions	Reach Steps
100-steps)	78.24 ± 1.86	$\mathbf{2.49 \pm 0.04}$	193.7 ± 3.2
RE(50-steps)	$\mathbf{81.90 \pm 2.81}$	2.75 ± 0.17	184.6 ± 1.4
Re(5-steps)	69.90 ± 7.34	5.23 ± 0.59	192.6 ± 3.3
Re(1-step)	59.57 ± 2.00	4.78 ± 0.16	$\mathbf{167.4 \pm 5.0}$
MC-based	74.14 ± 2.69	4.26 ± 0.56	192.8 ± 11.8

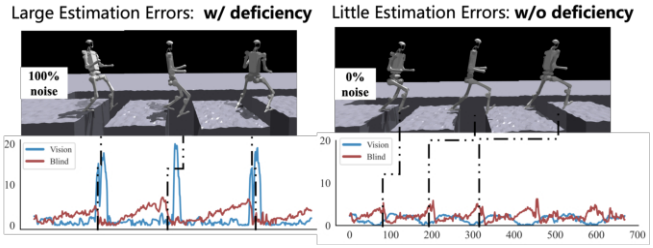

Fig. 8: The return estimation becomes noisy under perceptual deficiency.

We also evaluate the impact of different switch periods (T), which define the expected return duration during return estimator updates (Table V. While training performance remains consistent across varying periods, we observe that excessively short switch periods can negatively impact system performance. In such cases, the two policies may conflict, resulting in incomplete motion trajectories when traversing the challenging terrains and failures.

We observe that training effectiveness is highly dependent on data variance (Fig. 6-(d)). For instance, the estimator within vision policy converges the fastest due to its access to more accurate and comprehensive state observations, leading to fewer low-return instances. In contrast, the estimator within Noisy Perceptive and blind policies encounter more collisions and lower returns, causing their loss to degrade more slowly.

F. Hardware Performance

We make comparisons between VB-Com along with the vision policy and blind policy on G1 (Fig. 7-left), to demonstrate the superior performance of VB-Com in hardware

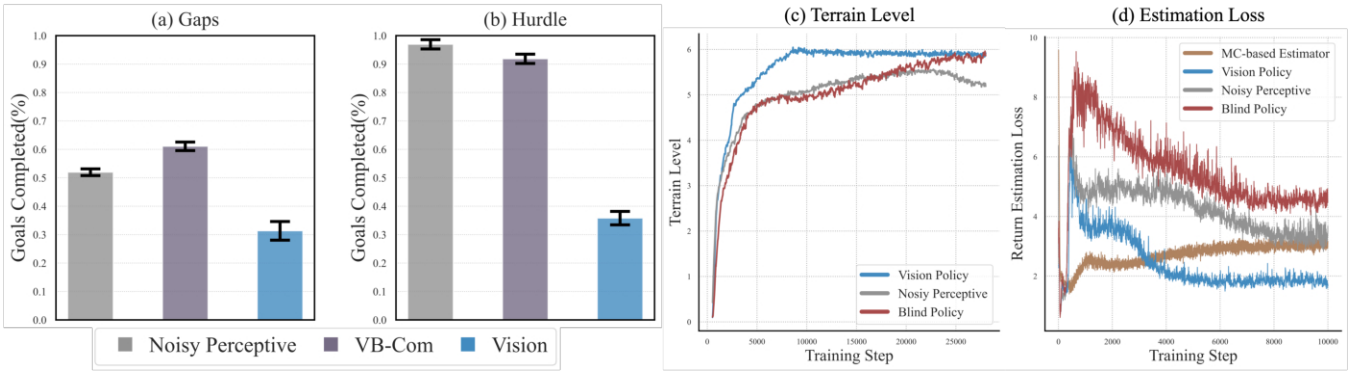

Fig. 6: (a,b) Comparisons between the Noisy Perceptive policy and VB-Com in navigating gaps and hurdles separately. (c,d) Training curves for terrain levels and the return estimation loss.

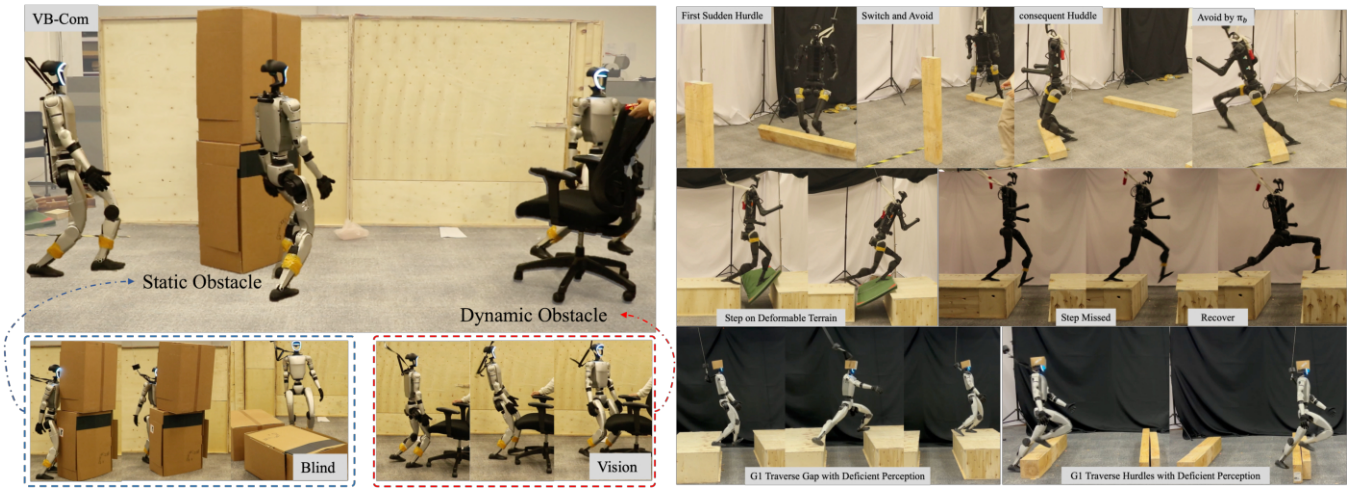

Fig. 7: (Left) Real-world comparisons of VB-Com, vision, and blind policies in obstacle avoidance on the G1. (Right)Hardware demonstrations on the robots traversing gaps and hurldes given deficient perception with VB-Com.

compared with signle policies. In the evaluation scenario, G1 encounters two consecutive obstacles along its path. The second dynamic obstacle obstructs the robot's direction before the elevation map can perceive it. VB-Com enables the robot to avoid the static obstacle without collision and subsequently avoid the dynamic obstacle after it collides with the suddenly appearing obstacle.

In contrast, for the baseline policies, the blind policy makes unnecessary contact with the static obstacles before avoiding them, which damages the environment. As for the vision policy, the robot collides with the obstacle and is unable to avoid it until the newly added obstacle is detected and integrated into the map.

Performance Against Deficient Perception. We demonstrate the ability of VB-Com to traverse challenging terrains given deficient perception (Fig. 7-right). We provide zero inputs for the heightmaps to evaluate the performance of VB-Com under perceptual deficiency. We introduce two consecutive hurdles, and the robot successfully recovers after colliding with them by switching to π_b. Additionally, we demonstrate that VB-Com enables recovery from a missed step on an unobserved gap. In this case, VB-Com saves the robot by performing a larger forward step to traverse the gap without perception, as the blind policy has learned during simulation.

VI. Conclusuins

VB-Com effectively integrates perceptual and proprioceptive observations to improve humanoid locomotion under conditions of perceptual degradation and dynamic disturbances. Our findings suggest that proprioceptive cues can successfully compensate for limited or noisy perceptual inputs.

Nonetheless, several limitations remain that open promising directions for future work. First, the current return estimator operates on scalar values, limiting its expressiveness in capturing the distinct capabilities of each sub-policy. A richer, structured representation could more accurately reflect the nuanced performance boundaries of individual policies. Second, while our dual-policy composition is effective for stable locomotion tasks, it could be generalized to a multi-policy selection framework to support broader operational scenarios. This would require more sophisticated switching mechanisms but could significantly enhance the flexibility and scalability of the system.

VII. ACKNOWLEDGMENT

This work is funded in part by the National Key R&D Program of China (2022ZD0160201), and Shanghai Artificial Intelligence Laboratory.

REFERENCES

[1] V. Makoviychuk et al., "Isaac gym: High performance gpu-based physics simulation for robot learning," *arXiv preprint arXiv:2108.10470*, 2021.

[2] K. Zakka et al., "Mujoco playground," *arXiv preprint arXiv:2502.08844*, 2025.

[3] G. B. Margolis and P. Agrawal, "Walk these ways: Tuning robot control for generalization with multiplicity of behavior," in *Conference on Robot Learning*, PMLR, 2023, pp. 22–31.

[4] J. Hwangbo et al., "Learning agile and dynamic motor skills for legged robots," *Science Robotics*, vol. 4, no. 26, eaau5872, 2019.

[5] I. Radosavovic, S. Kamat, T. Darrell, and J. Malik, "Learning humanoid locomotion over challenging terrain," *arXiv preprint arXiv:2410.03654*, 2024.

[6] Z. Chen et al., "Learning smooth humanoid locomotion through lipschitz-constrained policies," *arXiv preprint arXiv:2410.11825*, 2024.

[7] Z. Li, X. B. Peng, P. Abbeel, S. Levine, G. Berseth, and K. Sreenath, "Robust and versatile bipedal jumping control through reinforcement learning," *arXiv preprint arXiv:2302.09450*, 2023.

[8] J. Long et al., "Learning humanoid locomotion with perceptive internal model," *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[9] G. B. Margolis, G. Yang, K. Paigwar, T. Chen, and P. Agrawal, "Rapid locomotion via reinforcement learning," *The International Journal of Robotics Research*, vol. 43, no. 4, pp. 572–587, 2024.

[10] I. M. A. Nahrendra, B. Yu, and H. Myung, "Dreamwaq: Learning robust quadrupedal locomotion with implicit terrain imagination via deep reinforcement learning," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2023, pp. 5078–5084.

[11] J. Sun, L. Zhou, B. Geng, Y. Zhang, and Y. Li, "Leg state estimation for quadruped robot by using probabilistic model with proprioceptive feedback," *IEEE/ASME transactions on mechatronics*, 2024.

[12] J. Long, Z. Wang, Q. Li, L. Cao, J. Gao, and J. Pang, "Hybrid internal model: Learning agile legged locomotion with simulated robot response," in *The Twelfth International Conference on Learning Representations*, 2024.

[13] A. Kumar, Z. Fu, D. Pathak, and J. Malik, "Rma: Rapid motor adaptation for legged robots," *arXiv preprint arXiv:2107.04034*, 2021.

[14] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning quadrupedal locomotion over challenging terrain," *Science robotics*, vol. 5, no. 47, eabc5986, 2020.

[15] T. Miki, J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning robust perceptive locomotion for quadrupedal robots in the wild," *Science robotics*, vol. 7, no. 62, eabk2822, 2022.

[16] Z. Zhuang, S. Yao, and H. Zhao, "Humanoid parkour learning," *The Conference on Robot Learning (CoRL)*, 2024.

[17] R. Yang, M. Zhang, N. Hansen, H. Xu, and X. Wang, "Learning vision-guided quadrupedal locomotion end-to-end with cross-modal transformers," *arXiv preprint arXiv:2107.03996*, 2021.

[18] W. Cui et al., "Adapting humanoid locomotion over challenging terrain via two-phase training," in *8th Annual Conference on Robot Learning*.

[19] X. Cheng, K. Shi, A. Agarwal, and D. Pathak, "Extreme parkour with legged robots," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2024, pp. 11 443–11 450.

[20] A. Agarwal, A. Kumar, J. Malik, and D. Pathak, "Legged locomotion in challenging terrains using egocentric vision," in *Conference on robot learning*, PMLR, 2023, pp. 403–415.

[21] R. Yang, G. Yang, and X. Wang, "Neural volumetric memory for visual locomotion control," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1430–1440.

[22] D. Hoeller, N. Rudin, D. Sako, and M. Hutter, "Anymal parkour: Learning agile navigation for quadrupedal robots," *Science Robotics*, vol. 9, no. 88, eadi7566, 2024.

[23] S. Zhu, L. Mou, D. Li, B. Ye, R. Huang, and H. Zhao, "Vr-robo: A real-to-sim-to-real framework for visual robot navigation and locomotion," *arXiv preprint arXiv:2502.01536*, 2025.

[24] S. Choi et al., "Learning quadrupedal locomotion on deformable terrain," *Science Robotics*, vol. 8, no. 74, eade2256, 2023.

[25] C. Zhang et al., "Resilient legged local navigation: Learning to traverse with compromised perception end-to-end," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2024, pp. 34–41.

[26] Z. Fu, A. Kumar, A. Agarwal, H. Qi, J. Malik, and D. Pathak, "Coupling vision and proprioception for navigation of legged robots," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 273–17 283.

[27] X. Cheng, Y. Ji, J. Chen, R. Yang, G. Yang, and X. Wang, "Expressive whole-body control for humanoid robots," *arXiv preprint arXiv:2402.16796*, 2024.

[28] M. Ji et al., "Exbody2: Advanced expressive humanoid whole-body control," *arXiv preprint arXiv:2412.13196*, 2024.

[29] C. Lu et al., "Mobile-television: Predictive motion priors for humanoid whole-body control," *arXiv preprint arXiv:2412.07773*, 2024.

[30] T. He et al., "Hover: Versatile neural whole-body controller for humanoid robots," *arXiv preprint arXiv:2410.21229*, 2024.

[31] Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn, "Humanplus: Humanoid shadowing and imitation from humans," *arXiv preprint arXiv:2406.10454*, 2024.

[32] T. He et al., "Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning," *arXiv preprint arXiv:2406.08858*, 2024.

[33] X. Gu et al., "Advancing humanoid locomotion: Mastering challenging terrains with denoising world model learning," *arXiv preprint arXiv:2408.14472*, 2024.

[34] T. Miki, L. Wellhausen, R. Grandia, F. Jenelten, T. Homberger, and M. Hutter, "Elevation mapping for locomotion and navigation using gpu," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2022, pp. 2273–2280.

[35] R. Yu, Q. Wang, Y. Wang, Z. Wang, J. Wu, and Q. Zhu, "Walking with terrain reconstruction: Learning to traverse risky sparse footholds," *arXiv preprint arXiv:2409.15692*, 2024.

[36] J. Chen, J. Frey, R. Zhou, T. Miki, G. Martius, and M. Hutter, "Identifying terrain physical parameters from vision-towards physical-parameter-aware locomotion and navigation," *IEEE Robotics and Automation Letters*, 2024.

[37] J. Ren, Y. Liu, Y. Dai, J. Long, and G. Wang, "Top-nav: Legged navigation integrating terrain, obstacle and proprioception estimation," *arXiv preprint arXiv:2404.15256*, 2024.

[38] X. B. Peng, M. Chang, G. Zhang, P. Abbeel, and S. Levine, "Mcp: Learning composable hierarchical control with multiplicative compositional policies," *Advances in neural information processing systems*, vol. 32, 2019.

[39] A. Gupta, C. Lynch, B. Kinman, G. Peake, S. Levine, and K. Hausman, "Bootstrapped autonomous practicing via multi-task reinforcement learning," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2023, pp. 5020–5026.

[40] P.-L. Bacon, J. Harb, and D. Precup, "The option-critic architecture," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, 2017.

[41] D. Shah et al., "Value function spaces: Skill-centric state abstractions for long-horizon reasoning," *arXiv preprint arXiv:2111.03189*, 2021.

[42] S. Nasiriany, H. Liu, and Y. Zhu, "Augmenting reinforcement learning with behavior primitives for diverse manipulation tasks," in *2022 International Conference on Robotics and Automation (ICRA)*, IEEE, 2022, pp. 7477–7484.

[43] H. Zhang, W. Xu, and H. Yu, "Policy expansion for bridging offline-to-online reinforcement learning," *arXiv preprint arXiv:2302.00935*, 2023.

[44] Z. Zhuang et al., "Robot parkour learning," *arXiv preprint arXiv:2309.05665*, 2023.

[45] T. He, C. Zhang, W. Xiao, G. He, C. Liu, and G. Shi, "Agile but safe: Learning collision-free high-speed legged locomotion," *arXiv preprint arXiv:2401.17583*, 2024.

[46] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," *arXiv preprint arXiv:1506.02438*, 2015.