

REACH: An RL-based Hierarchical Humanoid Crawling Framework for Scalable Confined Space Traversal

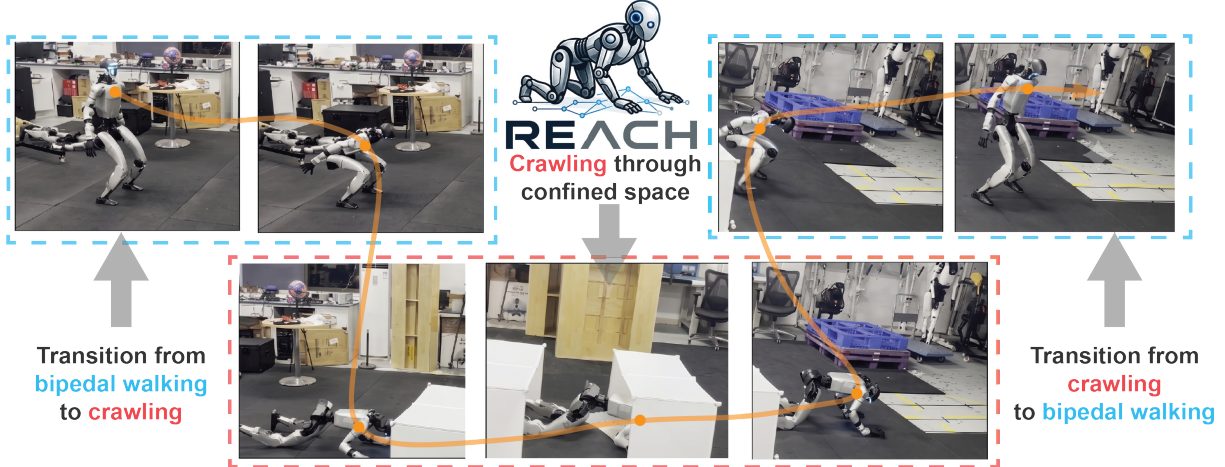


Fig. 1. We introduce REACH, a modular, scalable, RL-based hierarchical framework for humanoid crawling, enabling the Unitree G1 to transition smoothly between bipedal walking and crawling while traversing confined spaces.

Abstract—Humanoid locomotion in confined spaces requires continuous height adaptation across multiple locomotion modes, yet most existing approaches remain limited to height-adjustable bipedal walking. We present REACH, an RL-based hierarchical framework that enables a Unitree G1 humanoid to traverse height-constrained environments through variable-height bipedal walking and crawling. The learned policies regulate body height over complementary command ranges of 1.27–0.50 m in bipedal locomotion and 0.50–0.20 m in crawling, extending traversal beyond the clearance limits of crouched walking. To address contact-rich crawling dynamics, we employ a history-conditioned latent representation that captures terrain and interaction context from proprioceptive observations. A bidirectional transition policy further connects crouched bipedal locomotion and crawling across large kinematic and contact discontinuities. These skills are integrated with a LiDAR-based planner that converts point clouds into collision-free three-dimensional paths and locomotion commands, including proactive height targets. Real-world experiments demonstrate accurate height and velocity tracking, robust terrain traversal, reliable mode transitions and confined-space traversal with the proposed pipeline. Code and videos will be anonymously available at the [\[https://anonymous.4open.science/w/REACH/\]](https://anonymous.4open.science/w/REACH/).

I. INTRODUCTION

Confined-space traversal is essential to extending humanoid robots from structured settings to environments with narrow passages [1], [2]. However, the existing literature remains dominated by bipedal gaits [3], with torso height either fixed or adjusted only for crouched walking. This restriction prevents existing systems from accessing severely constrained spaces. In recent years, reinforcement learning (RL) has substantially enhanced humanoid mobility and robustness [4], [5], [6] or posture adaptation under complex

terrain and overhead constraints [7], [8], [9], [10]. Nevertheless, most approaches remain bipedal and are limited by leg workspace: aggressive crouching reduces maneuverability and balance margins.

However, extending a conventional humanoid from bipedal walking to another motion modality, e.g., crawling, introduces some challenges. First, the two modes differ substantially in configuration—directly switching between two independently trained policies requires the robot to pass through dynamically sensitive intermediate states and can produce discontinuous actions, imbalance, or falls. Second, crawling is intrinsically contact-rich: body segments such as the palms, wrists, knees, feet interact with the environment, generating impacts and disturbances that are difficult to model, leading to severe sim-to-real gaps.

Towards a practical solution for large-range height adaptation, robust multi-contact control, smooth transitions between locomotion modes, we present REACH (see Fig. 1), an RL-based hierarchical framework for scalable confined-space traversal using an unmodified Unitree G1 humanoid. Inspired by latent terrain-inference approaches such as DreamWaQ [11], we design a history-conditioned latent adaptation module tailored to contact-rich humanoid crawling. The module extracts terrain and interaction context from proprioceptive histories, improving robustness to multi-contact disturbances and real-world deployment conditions. A bidirectional motion-tracking policy connects crouched walking and crawling, enabling stable transitions across large kinematic and contact discontinuities. At the upper level, an adapted EGO-Planner [12] converts onboard LiDAR observations into collision-free 3D paths and proactive height commands. This hierarchical decomposition separates perception and

planning from locomotion, allowing integration with other navigation methods. Our main contributions are:

- A scalable locomotion framework extending locomotion beyond the clearance limits of crouched walking. The proposed framework combines variable-height bipedal walking (from 1.27 to 0.50 m) and crawling with height commands from 0.50 to 0.20 m (in experiments, real-world traversal of clearances below 0.40 m).
- A history-conditioned latent adaptation mechanism and a bidirectional transition policy for contact-rich crawling and bipedal–crawling switching.
- A practical planner–controller hierarchy that interfaces learned skills with LiDAR-based planning, validated successful real-world confined-space experiments.

II. RELATED WORK

A. Humanoid Locomotion under Complex Terrains

Recent learning-based humanoid locomotion methods increasingly incorporate exteroceptive perception to anticipate geometric constraints, using depth images [13], [14], LiDAR-derived elevation maps [15], [16], [17], or RGB observations [18]. These approaches improve mobility over rough, deformable, and discontinuous terrain, while model-informed objectives further enhance stability. Xie et al. [19] introduce a Zero-Moment Point-based reward for balance recovery, whereas Dai et al. [20] incorporate Control Lyapunov Function constraints for traversing sparse footholds. Attention-based terrain encoders additionally support accurate foothold selection [16], [17]. Related systems achieve robust blind locomotion through large-scale reinforcement learning [7] and terrain-aware traversal of stairs and gaps by combining online reconstruction with adaptive gait control [8]. However, these methods primarily assume vertically unconstrained environments and remain centered on bipedal locomotion, limiting their applicability to severely narrow passages.

B. Humanoid Postural Adaptation for Confined Spaces

Confined-space traversal requires substantial posture adaptation under low ceilings and overhanging obstacles. Perception-driven methods adapt humanoid posture online: HumanoidPF [9] enables crouching and obstacle negotiation, while Gallant [10] uses voxel representations and reinforcement learning under overhead constraints. Both remain predominantly bipedal, however, and aggressive crouching is limited by leg workspace, reduced maneuverability, and degraded disturbance rejection. Specialized low-profile platforms and hardware adaptations further ease confined locomotion [21], [3], but their morphological assumptions do not transfer directly to human-sized humanoids. Bridging large-range height adaptation with contact-rich crawling on a standard humanoid therefore remains an open challenge.

C. Motion Tracking and Multi-Contact Crawling

Motion tracking and imitation learning provide a complementary route to whole-body humanoid skills. BeyondMimic [22] demonstrates versatile control by tracking richly generated motions, while generative models such as Kimodo [23]

synthesize controllable whole-body references. We adopt this tracking paradigm only for bipedal–crawling mode switching: a single bidirectional tracker follows generated crawl-to-squat references (and their reverse), rather than building a general skill library. Separately, low-profile crawling has long been studied with model-based gait design [24], compliant or specialized hardware [3], [25], and morphologies that favor multi-contact support [21]. In contrast, our method learns command-conditioned crawling on an unmodified Unitree G1, uses history-conditioned latent adaptation for contact-rich disturbances, and exposes a modular interface to LiDAR-based planning.

III. METHODOLOGY

A. RL-based Variable-Height Locomotion

As shown in Fig. 2, we employ a shared architecture to train separate variable-height bipedal and crawling policies. The architecture combines an asymmetric actor–critic formulation with a history-conditioned context estimator, adopted from [11]. Both policies share the same deployable observation and network interface, enabling command execution across locomotion modes without privileged sensing on the physical robot. The actor maps the current proprioceptive observation, estimated base velocity, and latent context to a 29-dimensional joint action $\mathbf{a}_t \sim \pi_\phi(\mathbf{a}_t | \mathbf{o}_t, \tilde{\mathbf{v}}_t, \mathbf{z}_t)$, where the 97-dimensional observation is $\mathbf{o}_t = [\boldsymbol{\omega}_t, \mathbf{g}_t, \mathbf{c}_t, \boldsymbol{\theta}_t, \dot{\boldsymbol{\theta}}_t, \mathbf{a}_{t-1}]^T$. Here, $\mathbf{c}_t = [v_x^{\text{cmd}}, v_y^{\text{cmd}}, \omega_z^{\text{cmd}}, h^{\text{cmd}}]^T$ specifies planar velocity, yaw rate, and target height. A context estimator processes the five-step history $\mathbf{o}_t^H = [\mathbf{o}_t, \dots, \mathbf{o}_{t-4}]^T$. Its encoder produces the latent context $\mathbf{z}_t$ and estimated base velocity $\tilde{\mathbf{v}}_t$, while a decoder reconstructs the subsequent observation. Joint velocity-estimation and future-observation reconstruction objectives encourage $\mathbf{z}_t$ to encode terrain properties, contact events and dynamics that cannot be recovered from a single observation.

1) *Bipedal Walking*: For training, the critic receives $\mathbf{s}_t^{\text{walk}} = [\mathbf{o}_t, \mathbf{v}_t, \mathbf{h}_t^{\text{scan}}, \mathbf{p}_t^{\text{feet}}, \mathbf{v}_t^{\text{feet}}, \mathbf{f}_t^{\text{feet}}, h_t^{\text{root}}]^T$, which augments proprioception with terrain and foot-contact information available only in simulation.

2) *Crawling*: The crawling policy retains the same actor and context estimator, while the critic explicitly represents the four hand-and-knee contacts $\mathbf{s}_t^{\text{crawl}} = [\mathbf{o}_t, \mathbf{v}_t, \mathbf{h}_t^{\text{scan}}, \mathbf{p}_t^{\text{h\&k}}, \mathbf{v}_t^{\text{h\&k}}, \mathbf{f}_t^{\text{h\&k}}, h_t^{\text{root}}]^T$. This 306-dimensional input exposes the critic to contact-rich crawling dynamics, while the latent estimator helps the actor compensate for impacts, uncertain contacts, and modeling errors using onboard proprioception alone. Shared network interfaces therefore provide a common learning structure, whereas mode-specific privileged observations and rewards specialize training of the two locomotion policies.

3) *Reward Design*: Although the walking and crawling policies share the same command-conditioned learning architecture, they solve different contact-control problems. For mode $m \in \{\text{walk}, \text{crawl}\}$,

$$r_t^m = r_{\text{cmd},t}^m + r_{h,t}^m + r_{\text{post},t}^m + r_{\text{gait},t}^m + r_{\text{reg},t}^m, \quad (1)$$

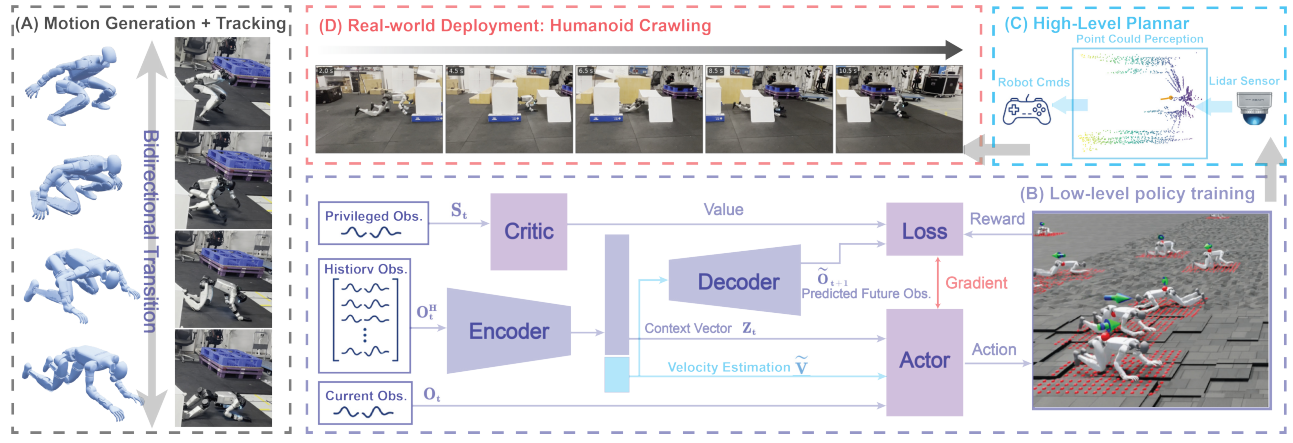


Fig. 2. Overview of REACH. (A) A bidirectional motion-tracking policy connects crouched bipedal locomotion and crawling. (B) The history-conditioned asymmetric actor-critic encodes proprioceptive histories, predicts future observations and body velocity, and uses privileged information only for critic training. (C) A model-based planner converts LiDAR point clouds into locomotion and height commands. (D) Real-world trials demonstrate crawling through consecutive low-clearance obstacles.

command tracking and regularization are shared, while height reference, posture prior, and gait shaping differ (see Appendix Tables II and III for full details of reward design). We highlight the two distinctions that most affect confined-space behavior. Let $\bar{h}^{\text{scan}} = \text{mean}(\mathbf{h}^{\text{scan}})$ and $\mathbf{c}_{\text{vel}}^{\text{cmd}} = [v_x^{\text{cmd}}, v_y^{\text{cmd}}, \omega_z^{\text{cmd}}]^\top$.

a) *Mode-specific height regulation.*: Both policies interpret h^{cmd} relative to local terrain, but track different body points:

$$r_h^{\text{walk}} = \exp\left[-25(h^{\text{head}} - \bar{h}^{\text{scan}} - h^{\text{cmd}})^2\right], \quad (2)$$

$$r_h^{\text{crawl}} = \exp\left[-\frac{(h^{\text{base}} - \bar{h}^{\text{scan}} - h^{\text{cmd}})^2}{0.15^2}\right]. \quad (3)$$

Walking uses head height as an overhead-clearance proxy, with torso-pitch constraints (see Table II) preventing trivial satisfaction by upper-body tilt. Crawling uses base height as a more stable clearance measure when the torso is nearly horizontal, and the head moves with arm placement.

b) *Phase-indexed walking versus phase-free crawling.*: Bipedal walking uses an explicit periodic contact schedule. For each foot $l \in \mathcal{L}_{\text{feet}}$,

$$\begin{aligned} \phi_t^l &= \left(\frac{t \bmod T}{T} + o^l\right) \bmod 1, \\ T &= 0.8 \text{ s}, \quad (o^{\text{LF}}, o^{\text{RF}}) = (0, 0.5), \end{aligned} \quad (4)$$

with desired stance $D_t^l = \mathbb{1}(\phi_t^l < 0.55)$ and measured contact $C_t^l = \mathbb{1}(F_{z,t}^l > 1)$. Matching these states yields

$$r_{\text{gait}}^{\text{walk}} = \mathbb{1}(\|\mathbf{c}_{\text{vel}}^{\text{cmd}}\| > 0.1) \frac{1}{2} \sum_{l \in \mathcal{L}_{\text{feet}}} \mathbb{1}(C_t^l = D_t^l), \quad (5)$$

encouraging alternating support with a duty factor close to normal walking [26]. A fixed four-limb phase clock is less suitable for crawling, where timing varies with height, terrain, and hand-knee impacts. Instead, for $\mathcal{L}_{\text{h\&k}} = \{\text{LH, RH, LK, RK}\}$ and $C_t^l = \mathbb{1}(F_{z,t}^l \geq 20)$, we penalize

unsafe support combinations

$$p_{\text{pair}} = \mathbb{1}(C^{\text{LH}} + C^{\text{RH}} = 0) + \mathbb{1}(C^{\text{LK}} + C^{\text{RK}} = 0), \quad (6)$$

$$p_{\text{ipsi}} = \mathbb{1}(C^{\text{LH}} + C^{\text{LK}} = 0) + \mathbb{1}(C^{\text{RH}} + C^{\text{RK}} = 0), \quad (7)$$

which preserve longitudinal and lateral support and favor diagonal hand-knee coordination without prescribing an exact stance sequence. Mode-duration and left-right timing terms further suppress contact chatter and asymmetric cycles:

$$r_{\text{mode}} = \mathbb{1}(\|\mathbf{c}_{\text{vel}}^{\text{cmd}}\| > 0.1) \min\left(\min_{l \in \mathcal{L}_{\text{h\&k}}} \Delta_t^l, 0.4\right), \quad (8)$$

$$p_{\text{sym}} = \frac{\mathbb{1}(\|\mathbf{c}_{\text{vel}}^{\text{cmd}}\| > 0.1)}{4} \sum_{q \in \{\text{H, K}\}} \sum_{\tau \in \{\text{lc, la}\}} |t_{\tau}^{\text{L}q} - t_{\tau}^{\text{R}q}|. \quad (9)$$

where Δ_t^l is the elapsed duration of the active contact or aerial mode of limb l , and $t_{\text{lc}}^l, t_{\text{la}}^l$ are the most recently completed contact and aerial durations. Together these terms define a *phase-free* crawling objective—stable support and balanced timing without a global gait clock—in contrast to the phase-indexed walking reward.

4) *Training Settings*: The two modes share the planar-velocity range but use different yaw-rate and height ranges:

$$\begin{aligned} \mathcal{C}_{\text{walk}} : \quad & v_x^{\text{cmd}} \in [-0.6, 0.8], \quad v_y^{\text{cmd}} \in [-0.6, 0.6], \\ & \omega_z^{\text{cmd}} \in [-1.0, 1.0], \quad h^{\text{cmd}} \in [0.50, 1.27], \end{aligned}$$

$$\begin{aligned} \mathcal{C}_{\text{crawl}} : \quad & v_x^{\text{cmd}} \in [-0.6, 0.8], \quad v_y^{\text{cmd}} \in [-0.6, 0.6], \\ & \omega_z^{\text{cmd}} \in [-0.75, 0.75], \quad h^{\text{cmd}} \in [0.20, 0.50]. \end{aligned}$$

Linear velocity, yaw rate, and height are expressed in m s^{-1} , rad s^{-1} , and meters, respectively. For both modes, velocity commands are resampled every 5–10 s and height commands every 2–10 s. Both policies use the same terrain curriculum over box arrays, random rough ground, undulating terrain, slopes, and platforms, with difficulty increasing gradually. The termination criteria remain mode-specific. Walking terminates upon torso-terrain contact ($F_z^{\text{torso}} > 1$), whereas

crawling terminates when the torso crosses the rollover criterion ($g_x^{\text{torso}} < 0$). Combined with the termination penalty, these criteria encourage fall avoidance in walking and stable four-contact support in crawling.

B. Bidirectional Bipedal–Crawling Transition

The bipedal and crawling policies are trained independently and therefore operate around different posture and contact distributions. Direct switching is unreliable because neither policy is trained to recover from the intermediate states between a deep squat and four-limb support. We connect the two skills using a generated reference motion and a single bidirectional tracking policy, as illustrated in Fig. 2(A). This decomposition makes the framework extensible: a new locomotion skill can be connected by generating a transition between compatible endpoint postures and training a tracker, without modifying the existing low-level policies.

1) *Reference Motion Generation*: We use Kimodo [23] to generate a coherent transition from quadrupedal crawling to a deep squat. Compared with collecting and retargeting motion-capture data, motion generation provides direct control over the terminal postures, allowing them to match the boundary states of the learned locomotion policies. In our implementation, the crawling endpoint corresponds to the stationary crawling policy at its maximum commanded height of 0.5 m, whereas the squat endpoint matches the stationary bipedal policy at its minimum height of 0.5 m.

The generated motion lasts 5 s at 30 Hz. Whole-body keyframes guide the sequence from the initial crawl, through a compact posture with the legs advanced beneath the body, to a single-foot–single-knee support state with both hands lifted, and finally to a stabilized deep squat, as shown in Fig. 2(A). The text prompt specifies the same progression: sequential leg advancement, arm retraction, and successive knee lift. These constraints preserve the intended contact order while leaving the generator free to synthesize the intervening whole-body motion. We generate the dynamically harder crawl-to-squat direction; the reverse sequence is used only as a tracking reference, while the physics-based policy learns a feasible closed-loop realization of both directions.

2) *Unified Transition Motion Tracking*: Confined-space traversal requires both crawl-to-squat and squat-to-crawl switching; because contact changes and balance constraints are direction-dependent, a one-way tracker is insufficient. We therefore train a *single* transition controller, extending BeyondMimic [22], to realize both directions; bidirectionality therefore refers to this shared policy, not to kinematic time-reversal alone. In practice, half of the parallel environments track the fixed reference in crawl-to-squat order and the remainder track its reverse, with the playback direction resampled every 2–6 s so that one policy covers both transitions and their intermediate states. We also extend the observation history from one to three steps, providing sufficient temporal context to distinguish rising toward the squat from descending toward four-limb support (Fig. 2(A)).

To smooth switching transients, the supervisor first brings the robot to the shared 0.5 m endpoint under the active

policy, then runs the closed-loop tracker along the reference, and finally transfers control to the target policy, rather than swapping policies through unsupported intermediate states. State and dynamics randomization, with permissive termination, further trains the tracker to absorb handoff mismatch and external disturbances. Under strong randomization, the standard tracking objective occasionally admits a kneeling local optimum near the squat endpoint. Strict termination can remove this behavior but substantially reduces useful training data. We instead introduce phase-gated shaping terms, summarized in Appendix Table IV. Let t_c denote the current reference-frame index. During the rising phase, head- and knee-height rewards promote completion of the deep squat, foot-level and non-foot-contact penalties stabilize the final support configuration, and hand-level shaping maintains reliable palm support before lift-off. These terms target the ambiguous contact changes without constraining the entire trajectory, preserving the diversity and robustness of the original motion-tracking objective.

IV. PERCEPTION AND PLANNING

Figure 2(C) summarizes the hierarchical navigation framework. The model-based front end converts LiDAR measurements into a collision-free three-dimensional body path, while a lightweight tracking interface converts this path into the velocity and height commands accepted by the learned locomotion policies. A locomotion supervisor then dispatches these commands to the bipedal or crawling policy and invokes the bidirectional transition policy when the required mode changes.

A. Spatiotemporal Point-Cloud Representation

Each LiDAR scan is first processed by Elevation Mapping [27] to estimate the local ground surface and separate obstacle returns from traversable terrain. A single scan, however, is incomplete because of the sensor’s limited field of view and self-occlusion by the robot. We therefore transform the current obstacle points into the world frame using the pose estimated by Point-LIO [28] and fuse them with its accumulated static map. The resulting obstacle cloud combines current measurements with previously observed geometry, retaining overhanging structures that cannot be represented by a ground-height map alone. This temporally persistent representation is supplied to the front-end A* search and the subsequent path optimizer.

B. Humanoid-Constrained Spatial Planning

We adapt EGO-Planner [12], originally developed for aerial robots, to produce geometric paths compatible with variable-height humanoid locomotion. Let $\mathbf{Q} = \{\mathbf{q}_i\}_{i=1}^{N_c}$ denote the path control points, where $\mathbf{q}_i = [x_i, y_i, z_i]^T$. The original optimization $J_{\text{orig}}(\mathbf{Q}) = \lambda_s J_s + \lambda_c J_c + \lambda_d J_d$ balances smoothness, collision avoidance, and dynamic feasibility. Because timing, velocity tracking, and dynamic balance are handled by the low-level policies, our planner optimizes a static spatial path rather than a dynamically

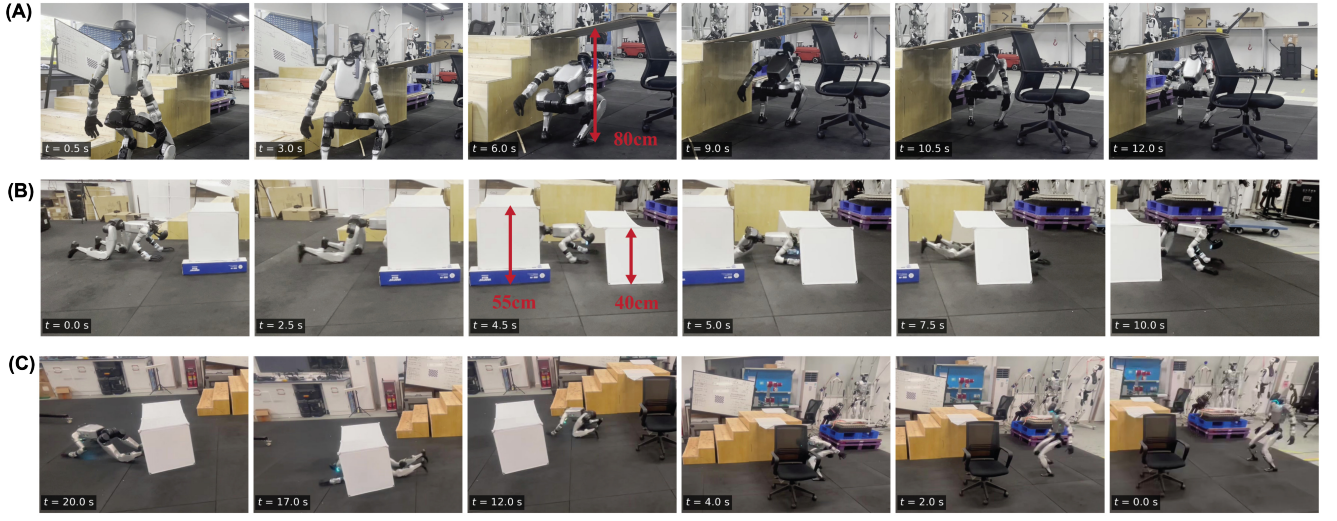


Fig. 3. Real-world evaluation of the complementary locomotion capabilities enabled by REACH. (A) Variable-height bipedal walking continuously adjusts body height to traverse an 80 cm-clearance obstacle. (B) Quadrupedal crawling enables robust multi-contact traversal through passages as low as 40 cm. (C) Bidirectional transitions between bipedal standing and quadrupedal crawling enable the robot to enter and exit low-clearance obstacles under different locomotion modes.

feasible aerial trajectory. We therefore remove J_d and introduce two humanoid-specific geometric terms: $J_{\text{ours}}(\mathbf{Q}) = \lambda_s J_s + \lambda_c J_c + \lambda_h J_h + \lambda_{ot} J_{ot}$.

a) Terrain-relative height regularization.: The first term regulates the body height along the path while allowing vertical adaptation when it is preferable to a long horizontal detour. Let $\hat{z}_g(x_i, y_i)$ be the estimated ground elevation beneath $\mathbf{q}_i$, and define the planned terrain-relative height as $h_i = z_i - \hat{z}_g(x_i, y_i)$. We use $J_h = \sum_{i=1}^{N_c} (h_i - h^{\text{nom}})^2$, where h^{nom} is the nominal walking height. In open space, this term suppresses unnecessary vertical oscillations. Near an overhead obstacle, its finite weight allows the optimizer to lower the path when the resulting height change is less costly than a lateral bypass. Importantly, expressing the cost relative to the local ground makes the planned height consistent with the height commands used to train both low-level policies.

b) Overhead traversal suppression.: An aerial planner may pass above an obstacle whenever doing so shortens the path, even when the obstacle exceeds the humanoid’s stepping capability. To suppress such fly-over solutions, we penalize path points above the obstacle footprint via $J_{ot} = \sum_{i \in \Omega_{\text{top}}} \exp(\kappa [z_i - z_{\text{obs}}(x_i, y_i)])$, where Ω_{top} contains control points whose horizontal coordinates lie within an obstacle footprint, $z_{\text{obs}}(x_i, y_i)$ is the obstacle’s top elevation, and $\kappa > 0$ controls the growth rate. The cost increases rapidly above the obstacle and thus acts as a soft vertical extension of its footprint. Together with J_c , it directs the optimizer around solid obstacles while preserving collision-free underpasses beneath elevated structures. The two added terms therefore distinguish feasible humanoid height adaptation from the unrestricted vertical motion assumed by the original aerial planner.

C. 3D Path-to-Policy Interface

The optimized path $\mathcal{P}$ is converted into low-dimensional commands by a three-dimensional Pure Pursuit tracker running at 30 Hz. At each update, a lookahead sphere of radius L_d , centered at the robot, selects a forward target on $\mathcal{P}$. After transformation into the robot base frame, this target is written as ${}^B \mathbf{p}_L = [x_L, y_L, z_L]^\top$. The horizontal components determine the commanded planar velocity, and the local path-heading error determines the yaw-rate command; velocity-coupled lateral feedback compensates for drift during path tracking. The vertical component is converted to a terrain-relative body-height command, providing advance notice of an approaching overhead constraint because the target lies ahead of the robot rather than at its current position. The resulting interface is $\mathbf{c}_t = [v_{x,t}^{\text{cmd}} \ v_{y,t}^{\text{cmd}} \ \omega_{z,t}^{\text{cmd}} \ h_t^{\text{cmd}}]^\top$, which exactly matches the command parameterization used during low-level policy training. The planned height also coordinates locomotion-mode selection. The bipedal and crawling policies meet at the shared switching height $h^{\text{sw}} = 0.5$ m: walking realizes commands above this boundary, whereas crawling realizes lower-clearance commands. When the requested height crosses the boundary, the supervisor first brings the active policy to the corresponding endpoint posture, executes the bidirectional transition policy, and then transfers control to the target locomotion policy. Commands are restricted to the training range of the active policy, preventing the planner from requesting heights or velocities outside its learned operating domain. Thus, a single geometric path jointly specifies where the robot should move, how its body height should change, and when a locomotion-mode transition is required.

V. EXPERIMENTS AND ANALYSIS

We organize the evaluation around three research questions: **RQ1**, how do the low-level ablations affect curriculum

TABLE I
ABLATION VARIANTS FOR DIFFERENT LOCOMOTION MODES.

Variant	Hist. estimator	Priv. scan	Idle penalty
FULL	✓	—	✓
NO-IDLE	✓	—	—
PROPRIO	—	—	✓
PRIV-SCAN	—	✓	✓

progression, which is an indicator of training progress; **RQ2**, how accurately do the resulting policies jointly track velocity and height commands; and **RQ3**, which training components enable robust bidirectional transitions and sim-to-real transfer? The first two questions evaluate variable-height walking and crawling in simulation. The third separates transition-policy ablations in simulation from deployment of the full policy on hardware. Complementing these controlled studies, Fig. 3 further shows **real-world scalability**: 80 cm-clearance walking, crawling as low as 40 cm, and composable skills in a single run that combines crouched walking and crawling via bidirectional transitions (Fig. 3(C)).

A. Variable-Height Locomotion

a) Ablation design.: For each locomotion mode, we train four policy variants in Isaac Lab [29] using 8192 parallel environments and 4000 training iterations. All variants receive current proprioception; Table I summarizes their differences. FULL is the proposed deployable architecture, in which the history-conditioned estimator provides the actor with an estimated body velocity $\tilde{v}_t$ and latent context z_t . NO-IDLE removes the linear- and angular-command idle penalties from FULL. PROPRIO removes both estimated quantities and uses only current proprioception. Finally, PRIV-SCAN augments current proprioception with privileged ground-truth terrain heights h^{scan} ; it is a simulation-only reference.

1) RQ1: Curriculum Progression.: Terrain difficulty increases only when the robot traverses a prescribed distance; its progression therefore measures how quickly a policy acquires command-responsive locomotion on increasingly complex terrain. Figure 4(a,b) reports the curriculum level throughout training. For walking, FULL is the first variant to exceed terrain level 4, doing so at approximately 1700 iterations. PROPRIO and PRIV-SCAN eventually reach similar levels but require substantially more training, whereas NO-IDLE remains below level 3 at the 4000-iteration budget. The crawling task provides a clearer separation: at iteration 4000, FULL reaches approximately level 3.6 and NO-IDLE reaches approximately level 3.3 after a pronounced delay, while PROPRIO and PRIV-SCAN do not progress beyond the initial level. These comparisons isolate two effects. The gap between FULL and NO-IDLE shows that idle penalties accelerate the acquisition of command-responsive motion. The failure of PROPRIO and PRIV-SCAN in crawling further indicates that explicit terrain heights alone do not replace the temporal context and body-velocity estimate required for contact-rich locomotion.

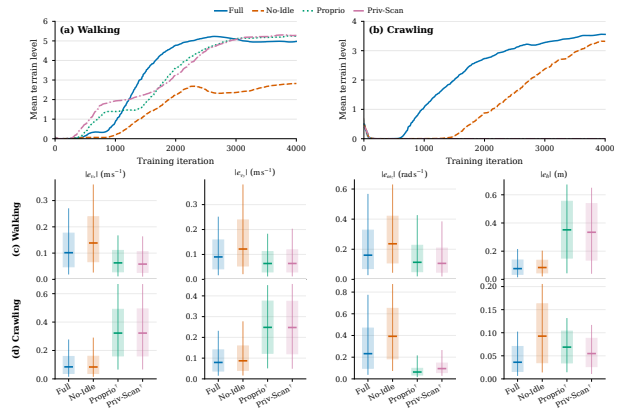


Fig. 4. Low-level locomotion ablations. (a,b) Mean terrain-curriculum level over training iterations. (c,d) Absolute command-tracking errors (median, IQR, 10th–90th percentiles; descriptive, not confidence intervals). †: crawling variants with no curriculum progress—low yaw-rate error should not be read as successful locomotion.

2) RQ2: Joint Command Tracking.: We evaluate each trained policy in 2048 parallel environments. A command is sampled from the corresponding training range at the beginning of each rollout and held fixed. Terrain levels are set to 4 for walking and 2 for crawling. After a 1-s warm-up, we record 9 s of absolute tracking error, $e_q(t) = |q(t) - q^{\text{cmd}}|$, $q \in \{v_x, v_y, \omega_z, h\}$, where h denotes the mode-specific terrain-relative height used by the corresponding low-level policy. Figure 4 jointly reports curriculum progression and the resulting tracking-error distributions. **In walking**, FULL reduces all three velocity errors relative to NO-IDLE. Although PROPRIO and PRIV-SCAN exhibit smaller velocity errors, their height errors are several times larger. Rollout inspection shows that these variants often remain near an upright posture instead of realizing the commanded crouched gait; their apparently favorable velocity errors therefore do not represent successful variable-height walking. FULL provides the best joint tracking of velocity and height. **In crawling**, FULL has forward-velocity error comparable to NO-IDLE, while reducing lateral, yaw-rate, and height errors. PROPRIO and PRIV-SCAN show smaller yaw-rate errors, but this must be interpreted together with Fig. 4(b): neither variant advances beyond the initial crawling terrain, and both have much larger translational errors. Their low yaw-rate errors therefore do not compensate for the absence of robust translational crawling. Across both evaluations, FULL is the only deployable variant that combines rapid curriculum progression with balanced tracking of all four commands. Because only one training run is available per condition, the curriculum curves support descriptive comparisons rather than cross-seed uncertainty estimates.

B. RQ3: Bidirectional Transition Robustness

a) Ablations and protocol.: We train three transition variants in Isaac Lab using 8192 environments and 4000 iterations. FULL (SIM.) uses the complete motion-tracking objective, phase-gated shaping terms, dynamics randomization, and permissive termination described in the methodol-

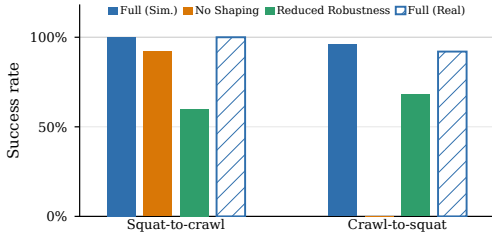


Fig. 5. Bidirectional transition success over 25 trials per direction. The first three configurations are evaluated in MuJoCo; the hatched FULL (REAL) bars denote deployment of the unchanged full policy on the Unitree G1.

ogy. NO SHAPING removes the additional phase-gated terms and retains only the BeyondMimic objective. REDUCED ROBUSTNESS halves all randomization magnitudes except friction and halves all termination thresholds except the rollover threshold, thereby combining weaker randomization with stricter termination. The three policies are evaluated in MuJoCo [30] for 25 trials in each direction. Every trial begins at the shared 0.5-m boundary posture. A squat-to-crawl trial is successful if the robot establishes stable hand-knee support without torso collision and subsequently moves under the crawling policy. Conversely, a crawl-to-squat trial must recover stable feet-only support without torso collision and subsequently walk under the bipedal policy. We additionally deploy the unchanged full policy on the physical Unitree G1 using the same initial states, criteria, and number of trials. This FULL (REAL) result is a hardware deployment of FULL (SIM.), not a fourth training configuration.

b) Results.: As shown in Fig. 5, FULL (SIM.) succeeds in 25/25 squat-to-crawl trials and 24/25 crawl-to-squat trials in MuJoCo, corresponding to success rates of 100% and 96%. NO SHAPING retains 23/25 squat-to-crawl successes but yields 0/25 crawl-to-squat successes. This asymmetric failure supports the intended role of the shaping terms: they are specifically necessary for escaping the kneeling optimum during crawl-to-squat recovery, rather than merely improving generic tracking accuracy. REDUCED ROBUSTNESS succeeds in only 15/25 and 17/25 trials, showing that broad randomization and permissive termination are important for recovery from switching-state variation. On hardware, FULL (REAL) achieves 25/25 squat-to-crawl and 23/25 crawl-to-squat successes, or 100% and 92%, respectively. The small decrease from simulation in the harder rising direction demonstrates that the learned bidirectional transition transfers to the physical robot while preserving reliable handover to both locomotion policies. Together with the qualitative real-world sequences in Fig. 3, these results indicate that the full REACH stack, including variable-height walking, contact-rich crawling, and mode switching, supports robust confined-space traversal on hardware.

VI. CONCLUSIONS

We presented REACH, an RL-based hierarchical framework that enables an unmodified Unitree G1 to traverse confined spaces through variable-height bipedal walking, crawling, bidirectional mode transitions, and a modular

LiDAR planner-controller interface. Ablation studies show that history-conditioned latent adaptation improves curriculum progression and joint command tracking, particularly for contact-rich crawling. The bidirectional transition policy transfers reliably from simulation to hardware, and real-world demos demonstrate complementary walking and crawling under low clearances with planner-issued velocity and height commands.

APPENDIX

(A) Reinforcement Learning Reward Summary

The signed coefficients in Tables II and III are applied directly to the listed nonnegative rewards or penalties. For compactness, define $\rho_{\mathcal{J}}(\theta^{\text{ref}}) = \sum_{i \in \mathcal{J}} |\theta_i - \theta_i^{\text{ref}}|$ and $\beta(\theta) = \frac{1}{2} [e^{-100|\min(\theta, 0)|} + e^{-100|\max(\theta - 1.2, 0)|}]$. Here, θ^{stand} and θ^{crawl} are the standing and crawling reference configurations, respectively; $\mathbf{g}_{xy}$ is the horizontal component of projected gravity; and τ_i^{PD} and τ_i are the PD-requested and post-saturation torques.

TABLE II

REWARD TERMS FOR VARIABLE-HEIGHT BIPEDAL WALKING.

Term	Definition	Weight
Head-height tracking	Eq. (2)	2.0
Yaw-rate tracking	$\bar{\omega}_{z,t} = \frac{1}{5} \sum_{k=0}^4 \omega_{z,t-k}; \quad r_{\omega} = \exp[-(\bar{\omega}_{z,t} - \omega_z^{\text{cmd}})^2 / 0.3^2]$	2.0
Linear idle	$\mathbb{1}(\ \mathbf{v}_{xy}^{\text{cmd}}\ > 0.2 \wedge \ \mathbf{v}_{xy}\ < 0.2)$	-3.0
Yaw idle	$\mathbb{1}[(\omega_z^{\text{cmd}} > 0.2 \wedge \omega_z < 0.1) \vee (\omega_z^{\text{cmd}} < -0.2 \wedge \omega_z > -0.1)]$	-2.0
Arm posture	$\rho_{\mathcal{J}_{\text{arm}}}(\theta^{\text{stand}})$	-0.2
Waist posture	$\rho_{\mathcal{J}_{\text{waist}}}(\theta^{\text{stand}})$	-1.0
Hip-roll posture	$\mathbb{1}(v_y^{\text{cmd}} \leq 0.2) \theta_{\text{hip roll}} - \theta_{\text{hip roll}}^{\text{stand}} $	-0.3
Torso-pitch range	$\beta(\theta_{\text{torso}}^{\text{pitch}})$	1.0
High-stance torso tilt	$\mathbb{1}(h^{\text{cmd}} > 1.0) \ \mathbf{g}_{xy}\ ^2$	-3.0
Phase-matched gait	Eq. (5)	1.0

TABLE III

REWARD TERMS FOR VARIABLE-HEIGHT HUMANOID CRAWLING.

Term	Definition	Weight
Base-height tracking	Eq. (3)	3.0
Waist-pitch posture	$ \theta_{\text{waist pitch}} - \theta_{\text{waist pitch}}^{\text{crawl}} $	-1.2
Wrist posture	$\rho_{\mathcal{J}_{\text{wrist}}}(\theta^{\text{crawl}})$	-0.2
Waist-roll posture	$\mathbb{1}(v_x^{\text{cmd}} < 0.3 \wedge \omega_z^{\text{cmd}} < 0.4) \theta_{\text{waist roll}} - \theta_{\text{waist roll}}^{\text{crawl}} $	-0.6
Hand/knee support	Eq. (6)	-0.6
Ipsilateral support	Eq. (7)	-0.3
Mode duration	Eq. (8)	1.0
Timing symmetry	Eq. (9)	-0.75
Torque saturation	$\sum_{i \in \mathcal{J}_{\text{all}}} \tau_i - \tau_i^{\text{PD}} $	-0.1

(B) Transition-Policy Reward Terms

Table IV lists the additional terms applied on top of the BeyondMimic objective. Define the filtered contact indicator $\chi_l(t) = \mathbb{1}(\max_{k \in \{0, \dots, 4\}} \|\mathbf{f}_{t-k, l}\|_2 > 1.0)$, where $\mathbf{f}_{t, l}$ is the contact force on link l . We use $\mathcal{L}$ for all links except the ankles, wrists, and knees, and $\mathcal{L}_{\text{nonfeet}}$ for all non-foot links.

TABLE IV

ADDITIONAL REWARD TERMS FOR BIDIRECTIONAL MOTION TRACKING.

Term	Definition	Weight
Head-height tracking	$\mathbb{1}(110 \leq t_c \leq 149) \exp[-(h^{\text{head}} - 0.5)^2 / 0.3^2]$	0.5
Knee-height tracking	$\mathbb{1}(110 \leq t_c \leq 149) \exp[-25(h^{\text{knees}} - 0.25)^2]$	1.0
Foot-level shaping	$\mathbb{1}(90 \leq t_c \leq 149) \ \mathbf{g}_{xy}^{\text{feet}}\ $	-0.4
Hand-level shaping	$\mathbb{1}(0 \leq t_c \leq 95) (\sqrt{1 - g_y^{\text{RH}}} + \sqrt{1 + g_y^{\text{LH}}})$	-0.6
Undesired contacts	$\sum_{l \in \mathcal{L}} \chi_l(t)$	-0.3
Squat-phase contacts	$\mathbb{1}(90 \leq t_c \leq 149) \sum_{l \in \mathcal{L}_{\text{nonfeet}}} \chi_l(t)$	-0.5

We used the following PPO parameters listed in Table V to train the policies for variable-height walking, crawling, and bidirectional transitions.

TABLE V

PPO HYPERPARAMETERS USED FOR WALKING, CRAWLING, AND TRACKING TRAINING.

Parameter	Value
Value loss coefficient	1.0
Value loss clip	0.2
Entropy coefficient	0.005
Epochs per iteration	5
Mini-batches per iteration	4
Simulation steps per iteration	20
Decimation	4
Parallel environments	8192
Learning rate	Adaptive
Discount factor γ	0.99
GAE parameter λ	0.95
Target KL divergence	0.01

REFERENCES

- [1] Z. Gu, J. Li, W. Shen, W. Yu, Z. Xie, S. McCrory, X. Cheng, A. Shamsah, R. Griffin, C. K. Liu, A. Kheddar, X. B. Peng, Y. Zhu, G. Shi, Q. Nguyen, G. Cheng, H. Gao, and Y. Zhao, "Humanoid locomotion and manipulation: Current progress and challenges in control, planning, and learning," *IEEE/ASME Transactions on Mechatronics*, vol. 31, no. 2, pp. 2300–2330, 2026.
- [2] T. Miki, J. Lee, L. Wellhausen, and M. Hutter, "Learning to walk in confined spaces using 3d representation," in *IEEE Int. Conference on Robotics and Automation (ICRA)*, 2024, pp. 8649–8656.
- [3] Z. Huang, X. Jiang, H. Liu, X. Chen, T. Fukuda, and Q. Huang, "Design of crawling motion for a biped walking humanoid with 3-dof rigid-flexible waist," in *2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*, 2018, pp. 974–979.
- [4] Z. Li, X. B. Peng, P. Abbeel, S. Levine, G. Berseth, and K. Sreenath, "Reinforcement learning for versatile, dynamic, and robust bipedal locomotion control," *The International Journal of Robotics Research*, vol. 44, no. 5, pp. 840–888, Apr. 2025.
- [5] Z. Li, X. Cheng, X. B. Peng, P. Abbeel, S. Levine, G. Berseth, and K. Sreenath, "Reinforcement learning for robust parameterized locomotion control of bipedal robots," in *IEEE Int. Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 2811–2817.
- [6] X. Ren, Y. Yang, Z. Weng, Z. Liu, and H. Kong, "Stubborn: A streamlined and unified reinforcement learning framework for robust motion tracking and fall recovery for humanoids," *arXiv:2606.12814*, 2026.
- [7] I. Radosavovic, S. Kamat, T. Darrell, and J. Malik, "Learning humanoid locomotion over challenging terrain," *arXiv preprint arXiv:2410.03654*, 2024.
- [8] H. Song, H. Zhu, T. Yu, Y. Liu, M. Yuan, W. Zhou, H. Chen, and H. Li, "Gait-adaptive perceptive humanoid locomotion with real-time under-base terrain reconstruction," *IEEE Robotics and Automation Letters*, 2025. [Online]. Available: <https://arxiv.org/abs/2512.07464>
- [9] H. Xue, S. Liang, Z. Zhang, Z. Zeng, Y. Liu, Y. Lian, J. Wang, Q. Liu, X. Shi, and L. Yi, "Collision-free humanoid traversal in cluttered indoor scenes," 2026. [Online]. Available: <https://arxiv.org/abs/2601.16035>
- [10] Q. Ben, B. Xu, K. Li, F. Jia, W. Zhang, J. Wang, J. Wang, D. Lin, and J. Pang, "Gallant: Voxel grid-based humanoid locomotion and local-navigation across 3d constrained terrains," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.
- [11] I. M. A. Nahrendra, B. Yu, and H. Myung, "Dreamwaq: Learning robust quadrupedal locomotion with implicit terrain imagination via deep reinforcement learning," in *IEEE Int. Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 5078–5084.
- [12] X. Zhou, Z. Wang, H. Ye, C. Xu, and F. Gao, "EGO-Planner: An ESDF-free gradient-based local planner for quadrotors," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 478–485, 2021.
- [13] S. Zhu, Z. Zhuang, M. Zhao, K.-Y. Lee, and H. Zhao, "Hiking in the wild: A scalable perceptive parkour framework for humanoids," 2026.
- [14] Y. Zhang, Y. Seo, J. Chen, Y. Yuan, K. Sreenath, P. Abbeel, C. Sferazza, K. Liu, R. Duan, and G. Shi, "Rpl: Learning robust humanoid perceptive locomotion on challenging terrains," *arXiv:2602.03002*.
- [15] J. Long, J. Ren, M. Shi, Z. Wang, T. Huang, P. Luo, and J. Pang, "Learning humanoid locomotion with perceptive internal model," in *IEEE Int. Conference on Robotics and Automation (ICRA)*, 2025.
- [16] J. He, C. Zhang, F. Jenelten, R. Grandia, M. Bächer, and M. Hutter, "Attention-based map encoding for learning generalized legged locomotion," *Science Robotics*, 2025.
- [17] C. Zhang, V. Klemm, F. Yang, and M. Hutter, "AME-2: Agile and generalized legged locomotion via attention-based neural map encoding," 2026. [Online]. Available: <https://arxiv.org/abs/2601.08485>
- [18] Y. Liu, X. Chen, H. Pan, C. Shi, Y. Chen, K. Huang, Z. Zeng, and H. Lu, "Geoloco: Leveraging 3d geometric priors from visual foundation model for robust rgb-only humanoid locomotion," *arXiv:2603.07624*, 2026.
- [19] W. Xie, C. Bai, J. Shi, J. Yang, Y. Ge, and W. Zhang, "Humanoid whole-body locomotion on narrow terrain via dynamic balance and reinforcement learning," in *IEEE/RSJ Int. Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025.
- [20] M. Dai, W. D. Compton, J. Li, L. Yang, and A. D. Ames, "Walk the planc: Physics-guided rl for agile humanoid locomotion on constrained footholds," in *IEEE/RSJ Int. Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2026.
- [21] A. Schperberg, Y. Tanaka, S. Di Cairano, and D. Hong, "Mobius: A multi-modal bipedal robot that can walk, crawl, climb, and roll," in *Robotics: Science and Systems (RSS)*, 2026.
- [22] Q. Liao, T. E. Truong, X. Huang, Y. Gao, G. Tevet, K. Sreenath, and C. K. Liu, "BeyondMimic: From motion tracking to versatile humanoid control via guided diffusion," *arXiv:2508.08241*, 2025.
- [23] D. Rempe, M. Petrovich, Y. Yuan, H. Zhang, X. B. Peng, Y. Jiang, T. Wang, U. Iqbal, D. Minor, M. de Ruyter, J. Li, C. Tessler, E. Lim, E. Jeong, S. Wu, E. Hassani, M. Huang, J.-B. Yu, C. Chung, L. Song, O. Dionne, J. Kautz, S. Yuen, and S. Fidler, "Kimodo: Scaling controllable human motion generation," *arXiv:2603.15546*, 2026.
- [24] G. Brooks, P. Krishnamurthy, and F. Khorrami, "Low-profile crawling for humanoid motion in tight spaces," in *IEEE/RSJ Int. Conference on Intelligent Robots and Systems (IROS)*, 2015.
- [25] J. Ma, Y. Dong, W. Li, and P. Lu, "MASH: MAstering safe and high-speed quadruped navigation with adaptive gait transitions," *IEEE Transactions on Field Robotics*, June 2026.
- [26] D. A. Winter, "Kinematic and kinetic patterns in human gait: Variability and compensating effects," *Human Movement Science*, vol. 3, no. 1–2, pp. 51–76, 1984.
- [27] P. Fankhauser, M. Bloesch, and M. Hutter, "Probabilistic terrain mapping for mobile robots with uncertain localization," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3019–3026, 2018.
- [28] C. Zheng *et al.*, "Point-LIO: Robust high-bandwidth light detection and ranging inertial odometry," *IEEE Transactions on Robotics*, 2022.
- [29] M. Mittal, P. Roth, J. Tighe, A. Richard *et al.*, "Isaac lab: A gpu-accelerated simulation framework for multi-modal robot learning," *arXiv preprint arXiv:2511.04831*, 2025.
- [30] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *IEEE/RSJ Int. Conference on Intelligent Robots and Systems (IROS)*, 2012, pp. 5026–5033.