PGMT: Perceptive General Motion Tracking for Humanoid Robots

Hongyi LI[1], LI Peizhuo[2], Yucheng TAO[1], Ze WANG[1], Fangzhou XU[2], Jinyi CHEN[1], Yanyan YUAN[3], Dapeng JIA[2], Yongbin JIN[3,†], Mingfeng FAN[2,†], Guillaume SARTORETTI[2] and Hongtao WANG[1,3,‡]

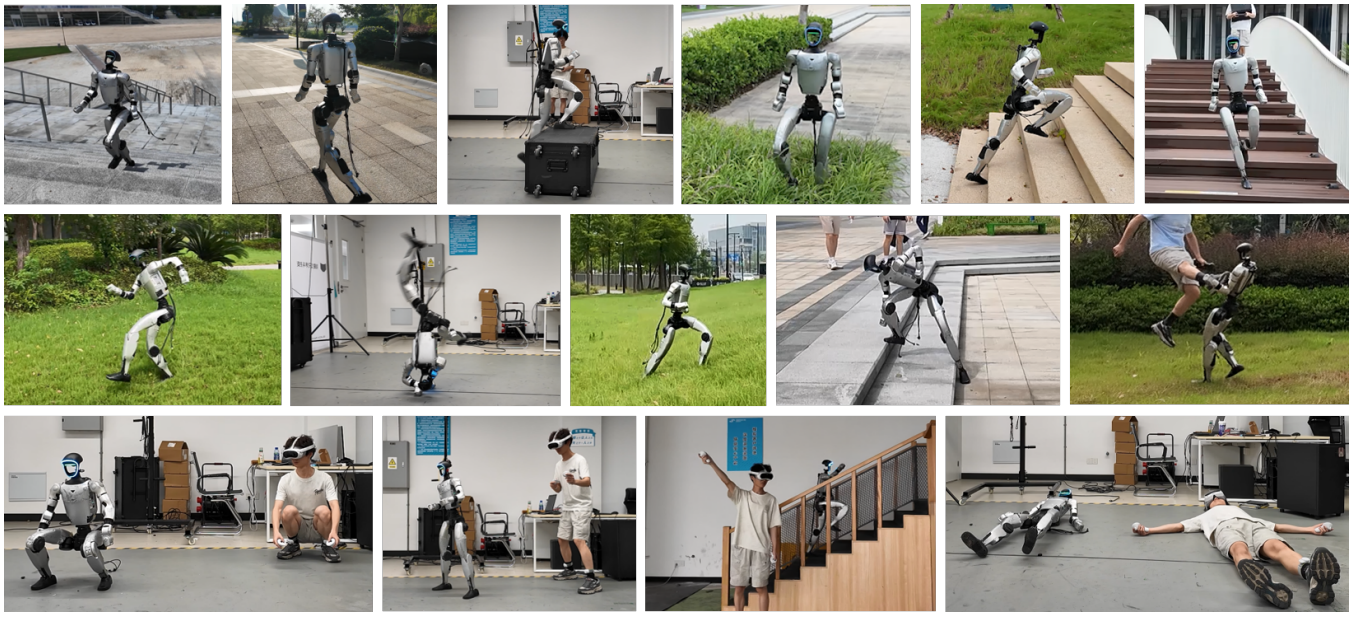

Fig. 1: **PGMT robustly executes whole-body motions across diverse real-world environments**, from structured indoor settings to stairs, uneven outdoor terrain, and dense vegetation. The same policy accommodates motion references from multiple command sources, providing a unified interface for perceptive whole-body locomotion and tracking.

***Abstract*— Humanoid motion trackers can reproduce diverse whole-body motions, but their performance degrades on complex terrain where terrain-agnostic references become physically infeasible. We present PGMT, a Perceptive General Motion Tracking pipeline for humanoid robots that learns terrain adaptation from independently selected motion references and terrains. PGMT first learns a general tracking and recovery prior, then incorporates terrain perception through motion-conditioned terrain glimpses that selectively encode regions relevant to the current motion. Terrain-aware tracking relaxation allows necessary deviations from the reference while preserving its motion intent. Zero-shot deployment on a Unitree G1 demonstrates robust terrain-adaptive locomotion and whole-body motion execution over real-world terrain with obstacles up to 37 cm high, while supporting teleoperation, dynamic motion tracking, and fall recovery. PGMT extends general humanoid motion tracking beyond flat ground, providing a unified policy for terrain-adaptive locomotion, diverse whole-body behaviors, and teleoperation in complex environments.**

I. Introduction

Recent advances in humanoid control have produced highly capable whole-body motion trackers that can reproduce diverse human motions with remarkable fidelity [1], [2]. Yet their success remains largely confined to flat ground. In complex environments, terrain-agnostic references may become physically infeasible, requiring the robot to actively adapt its contacts, swing trajectories, and whole-body posture according to local terrain observations rather than strictly reproduce the commanded motion [3], [4].

Such perceptive adaptation is particularly important for applications such as remote teleoperation and data collection, search and rescue over rough terrain, and transportation through complex environments [5]. In these scenarios, motion commands from human operators or higher-level planners cannot be expected to anticipate the exact geometry encountered by the robot. A general motion tracker should therefore use onboard perception to resolve local terrain constraints autonomously: the reference specifies *what motion to perform*, while environmental observations determine *how that motion should be physically realized*.

Learning such a capability, however, is difficult because large-scale motion and terrain data are naturally unpaired. Existing human motion datasets provide diverse whole-body behaviors but generally lack synchronized terrain geometry and terrain-specific adaptations [6], while collecting paired motion–terrain demonstrations becomes increasingly difficult as motion and terrain diversity grow. Recent work addresses this problem by constructing terrain-adapted references be-

[†]Corresponding Author [‡]Project Lead
[1]Center of X-Mechanics, Zhejiang University, Hangzhou, China.
[2]MARMot Lab, National University of Singapore, Singapore
[3]MirrorMe Technology Co., Ltd.

fore tracking [7]. Although effective, these approaches depend on the quality and generalization of the intermediate stage, while the need to construct terrain-adapted references limits their scalability to large-scale motion and terrain data. A general perceptive tracker should instead learn to relate motion intent directly to terrain constraints from independently available motion and terrain data.

We instead treat terrain adaptation as a capability of the tracking policy itself. The raw reference specifies the intended motion, while terrain perception provides the local physical constraints. By jointly processing both signals at the same level in an Intent Fusion Module (IFM), the policy can deviate from the reference when necessary to maintain feasibility, without requiring an upstream terrain-matched reference generator. Building on this formulation, we propose *PGMT*, a Perceptive General Motion Tracking pipeline that takes raw reference motions and local terrain observations directly as input. PGMT learns to attend to terrain regions relevant to the current motion and to adjust footholds and whole-body posture when the environment makes strict reference tracking infeasible, while preserving the underlying motion intent.

This formulation turns perceptive tracking into a reusable low-level interface between terrain-agnostic motion commands and terrain-aware physical execution. The same PGMT policy can execute dynamic whole-body motions over stairs, slopes, and uneven ground, serve as a perceptive locomotion controller, and support teleoperation in which the robot resolves local terrain constraints on the operator's behalf. More broadly, motion references from human operators, motion-matching systems, or higher-level planners can specify desired behavior without explicitly encoding the geometry under the robot's feet, while PGMT grounds these commands into physically feasible whole-body motion. The contributions of this work are as follows.

- We propose **PGMT**, a general motion tracker with perception that directly tracks terrain-agnostic motion references over complex terrain without requiring terrain-matched reference trajectories.
- We design a *progressively extensible architecture* that decouples motion and perception encoding, enabling new perceptual modalities to be injected while preserving the learned motion tracking capability.
- We conduct extensive simulations and real-world experiments across diverse terrains and motion-reference command sources, demonstrating the robust zero-shot generalization of PGMT.

II. Related Works

Large-Scale Motion Priors for Humanoid Control. General humanoid motion tracking grew out of large-scale retargeting and privileged policy distillation, enabling expressive real-world control through flexible command interfaces [2], [8], [9]. As the motion repertoire expanded, adaptive sampling, specialized experts, and latent representations addressed uneven motion difficulty and limited policy capacity [10], [11]. Embodiment mismatch can be handled by adapting motion before policy learning or transferring a tracker after training [12], [13]. Recovery training and dynamics adaptation instead target failures during deployment [14], [15]. Scaling data and compute broadens behavior coverage, while masked conditioning supports varied command formats [16], [17]. Even so, tracking depends on reference quality because retargeting artifacts and infeasible poses are passed to the policy [18]. The references also lack the geometry encountered at deployment and may conflict with terrain or contact constraints.

Perception-Aware Humanoid Control. Perception changes the problem from reproducing a reference to deciding how it should be executed on the observed terrain. Early learning-based approaches demonstrated whole-body obstacle crossing and used teacher–student training to cope with noisy terrain estimates [19], [20]. Later work focuses on both the reliability and relevance of perceptual input. Depth synthesis and global–local reasoning improve terrain representations [21], [22], whereas elevation maps and active gaze focus control on regions that matter for the current motion [23], [24]. In broader legged locomotion, separate proprioceptive and visuospatial pathways allow stable control when vision fails [25], and perception has also been connected with longer-horizon navigation [26]. These methods adapt well to terrain, but their objectives remain tied to particular locomotion or obstacle-crossing tasks rather than diverse whole-body references.

Perceptive Motion Priors and Tracking. Combining motion priors with perception exposes a data problem: large motion collections are rarely paired with terrain-specific adaptations. Existing work differs mainly in where it adds this missing information. Some methods construct supervision before policy learning from terrain-conformal references, scene geometry, motion-matched experts, or inferred contacts [3], [4], [27], [28]. Others adapt within a generation-tracking loop, using execution feedback to reshape references or fine-tune the tracker [7], [29]. Guided diffusion provides another route by building behaviors from motion primitives collected by a real-robot tracker [30]. These strategies avoid direct motion–terrain pairing, but still rely on intermediate signals whose errors propagate to the tracker.

III. Problem Formulation

We formulate perceptive whole-body motion tracking as a partially observable Markov decision process (POMDP). At each time step t, the policy p_θ takes an observation as input and outputs the target joint positions $\boldsymbol{a_t} \in \mathbb{R}^{29}$ as the action. The full perceptive observation is given by $\boldsymbol{y}_t = \{\boldsymbol{o}_t, \boldsymbol{H}_t, \boldsymbol{C}_t^K, \boldsymbol{M}_t\}$, where $\boldsymbol{o}_t \in \mathbb{R}^{96}$ denotes the current proprioception for immediate feedback, $\boldsymbol{H}_t \in \mathbb{R}^{10\times 96}$ represents a ten-frame proprioception history, $\boldsymbol{C}_t^K$ denotes future reference frames, and $\boldsymbol{M_t}$ represents the elevation map. Each frame in $\boldsymbol{o}_t$ and $\boldsymbol{H}_t$ contains reference-relative anchor orientation $\boldsymbol{e}_t \in \mathbb{R}^6$, base angular velocity $\boldsymbol{\omega}_t \in \mathbb{R}^3$, joint positions $\boldsymbol{q}_t \in \mathbb{R}^{29}$ and velocities $\dot{\boldsymbol{q}}_t \in \mathbb{R}^{29}$, and the previous action $\boldsymbol{a}_{t-1}$. We sample the future reference frames

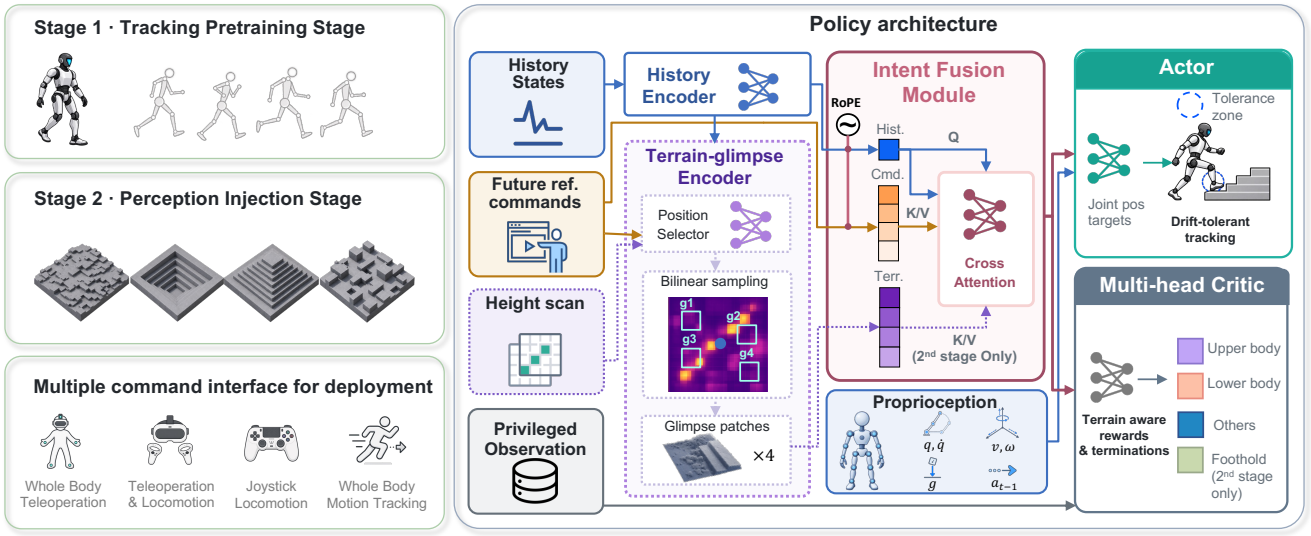

Fig. 2: **Overview of PGMT.** PGMT is trained in two stages: a Tracking Pretraining Stage that establishes a general whole-body motion tracking and recovery prior, followed by a Perception Injection Stage that selectively fuses local geometry with motion intent, enabling terrain adaptation without terrain-matched reference trajectories. The resulting unified policy supports terrain-adaptive locomotion, whole-body motion tracking, and multiple teleoperation interfaces.

at exponentially increasing temporal offsets:

$$\boldsymbol{C}_t^K = ([\boldsymbol{q}_\tau^r, \dot{\boldsymbol{q}}_\tau^r, \widetilde{\boldsymbol{v}}_\tau^r])_{k=0}^{K-1} \in \mathbb{R}^{K\times 61}, \tag{1}$$

where k indexes the future frames and $\tau = t + 2^k - 1$. The superscript r denotes reference quantities. Specifically, $\boldsymbol{q}^r \in \mathbb{R}^{29}$ and $\dot{\boldsymbol{q}}^r \in \mathbb{R}^{29}$ denote reference joint positions and velocities, respectively, while $\widetilde{\boldsymbol{v}}^r = (\widetilde{v}_x, \widetilde{v}_y, 0)$ denotes the corrected planar anchor-velocity command. Collectively, $\boldsymbol{C}_t^K$ provides the future motion intent to the policy. The elevation map $\boldsymbol{M}_t \in \mathbb{R}^{21\times 21}$ covers a yaw-aligned $2\,\mathrm{m} \times 2\,\mathrm{m}$ region and is introduced during the second training stage.

IV. Perceptive Generalized Motion Tracking

As illustrated in Fig. 2, PGMT is trained in two stages. The *Tracking Pretraining Stage* first learns a general terrain-agnostic prior for whole-body motion tracking, while the *Perception Injection Stage* subsequently incorporates terrain perception to enable terrain-aware adaptation without relearning the underlying tracking capability from scratch. In the first stage, the policy p_θ receives the non-perceptive observation $\boldsymbol{y}_t^{pre} = \{\boldsymbol{o}_t, \boldsymbol{H}_t, \boldsymbol{C}_t^K\}$ as input. In the second stage, the elevation map $\boldsymbol{M}_t$ is additionally introduced, yielding the full perceptive observation $\boldsymbol{y}_t^{\text{inj}} = \{\boldsymbol{o}_t, \boldsymbol{H}_t, \boldsymbol{C}_t^K, \boldsymbol{M}_t\}$.

A. Tracking Pretraining Stage

The first stage of PGMT learns a general whole-body motion tracking prior on flat terrain, providing a stable initialization for the subsequent injection of terrain perception. This stage adopts a terrain-agnostic policy architecture mainly consisting of an Intent Fusion Module (IFM), an actor, and a multi-head critic. The IFM integrates proprioceptions and future reference frames into a state-conditioned representation, which is used by the actor for low-level motion control and by the critic for value estimation. To train this stage, we combine a body-part-based reward decomposition with a recovery curriculum and adaptive motion sampling, enabling a unified policy to learn diverse whole-body motions.

1) Terrain-agnostic Policy Architecture: Given the current proprioceptive observation $\boldsymbol{o}_t$ and the proprioceptive history $\boldsymbol{H}_t$, we use $\boldsymbol{o}_t$ as the query to attend to the historical observations in $\boldsymbol{H}_t$, producing a compact history-conditioned state token:

$$\boldsymbol{s}_t^{\text{hist}} = \text{CrossAttn}_H\left(Q = \boldsymbol{o}_t,\, K = V = \boldsymbol{H}_t\right), \tag{2}$$

where CrossAttn. denotes a multi-head cross-attention (MHCA) layer, and Q, K, and V denote the query, key, and value, respectively. The resulting token $\boldsymbol{s}_t^{\text{hist}}$ provides a compact representation of recent robot dynamics and temporal context that is not available from the instantaneous proprioceptive observation alone.

Intent Fusion Module. To effectively fuse proprioceptive information with future motion references, we design the IFM consisting of an MHCA layer equipped with Rotary Position Embedding (RoPE). Specifically, the module uses the history-conditioned state token $\boldsymbol{s}_t^{\text{hist}}$ as the query to attend to the concatenation of the current state token and future reference frames $\boldsymbol{C}_t^K$, producing a state-conditioned motion-intent token:

$$\boldsymbol{s}_t^{\text{int}} = \text{CrossAttn}_C\left(Q = \boldsymbol{s}_t^{\text{hist}},\, K = V = [\boldsymbol{s}_t^{\text{hist}}; \boldsymbol{C}_t^K]\right). \tag{3}$$

This design allows the policy to selectively emphasize future reference information conditioned on the robot's current state and recent dynamics. The resulting intent token $\boldsymbol{s}_t^{\text{int}}$ is shared by the actor and critic and is further augmented with terrain information in the second training stage (Sec. IV-B).

Actor Network. The actor concatenates the current proprioceptive $\boldsymbol{o}_t$ with the motion-intent token $\boldsymbol{s}_t^{\text{int}}$ and maps the resulting representation to the control action:

$$\boldsymbol{a}_t = \text{MLP}_A\left([\boldsymbol{o}_t, \boldsymbol{s}_t^{\text{int}}]\right), \tag{4}$$

where MLP is a multi-layer perceptron. While s_t^{int} provides a compact representation of the state-conditioned motion intent, the direct inclusion of o_t preserves instantaneous proprioceptive feedback for fine-grained low-level control. The target joint positions are converted into joint torques by a low-level PD controller. The actor relies exclusively on information available at deployment time.

Multi-Head Critic. Whole-body motion tracking combines objectives with distinct learning characteristics. Upper-body objectives primarily preserve motion expression and pose fidelity, whereas lower-body objectives are closely coupled with balance and contact transitions. The remaining objectives shape root motion, recovery, safety, and control regularization. As summarized in Table I, we partition the pretraining reward into three groups:

$$r_t = r_t^{\text{upper}} + r_t^{\text{lower}} + r_t^{\text{aux}}, \tag{5}$$

where r_t^{upper} and r_t^{lower} contain the upper- and lower-body tracking objectives, respectively, and r_t^{aux} contains the remaining auxiliary objectives. A single scalar critic may entangle reward components with substantially different scales and learning dynamics. We therefore employ a multi-head critic that outputs a vector of value estimates $\boldsymbol{V}_t$, one per reward group:

$$\boldsymbol{V}_t = \text{MLP}_V\left([\boldsymbol{o}_t^{\text{priv}}, \boldsymbol{s}_t^{\text{int}}]\right) = \left[V_t^{\text{upper}}, V_t^{\text{lower}}, V_t^{\text{aux}}\right], \tag{6}$$

where $\boldsymbol{o}_t^{\text{priv}}$ denotes privileged observations available only during simulation training. The three heads share a common critic backbone, enabling representation sharing while modeling the reward groups separately.

2) Training Optimization: In this stage, we train PGMT using motions from LAFAN1 [31] retargeted to a humanoid robot. Motion segments are sampled across reference sequences and starting frames to train a unified policy over diverse whole-body behaviors. All pretraining is conducted on flat terrain. We adopt a root-centric tracking formulation, where the primary body-tracking errors are computed relative to the reference root anchor. In contrast to tracking full trajectories in the world frame, this formulation emphasizes relative body motion and reduces dependence on absolute global position and heading. Consequently, the learned tracking prior can reproduce the same motion intent from different initial poses and headings, providing a transferable initialization for subsequent terrain adaptation.

Recovery Curriculum and Adaptive Sampling. Following RGMT [14], we incorporate the recovery curriculum into this stage's training. We construct an offline pool of post-fall robot states and progressively increase the proportion of environments initialized from this pool according to episode-survival performance. In parallel, we employ adaptive sampling that assigns higher probabilities to motion segments associated with frequent tracking failures while retaining uniform coverage of the full motion dataset.

B. Perception Injection Stage

At this stage, we progressively extend the PGMT policy architecture through a perception injection process, enabling PGMT to adapt to diverse terrains.

TABLE I: Reward groups for PGMT.

Reward term	Weight	Reward term	Weight
The 1st & 2nd Stages			
Upper body			
Link position	1.0	Link orientation	1.0
Link linear velocity	0.5	Link angular velocity	0.5
Joint position	1.0	Joint velocity	0.5
Lower body			
TA link position	0.5	TA link orientation	2.0
Link linear velocity	0.5	Link angular velocity	0.5
TA joint position	0.5	Joint velocity	0.5
Auxiliary			
Root orientation	0.5	Corrected root velocity	2.0
Floating-anchor position	1.0	Recovery upward velocity	12.5
Pelvis vertical acceleration	-10^{-3}	EE acceleration mismatch	-10^{-3}
Action rate	-0.05	Joint limit	-15.0
Undesired contact	-0.1	Head/torso impact	-10^{-5}
The 2nd Stage			
Terrain-contact			
Touchdown quality	10.0	Reference contact match	1.5
Slip	-1.0	Stumble	-20.0
Contact switching	-30.0	Contact force	-10^{-6}

1) Perception Injection Process: Building upon the terrain-agnostic policy architecture, we introduce a Terrain-Glimpse Encoder that generates state-conditioned terrain tokens, which are subsequently integrated into the IFM to facilitate terrain-aware low-level motion control. In addition, we extend the multi-head critic to evaluate a newly introduced terrain-contact objective.

Terrain-glimpse Encoder. Directly encoding the entire elevation map as a dense terrain representation for the actor is computationally expensive and may introduce unnecessary noise, as only a small subset of terrain regions is typically relevant to locomotion, particularly those around potential footholds and terrain boundaries. Instead, PGMT employs a Terrain-Glimpse Encoder that leverages the elevation map to localize a sparse set of locomotion-relevant regions conditioned on the robot's dynamics and anticipated motions. These regions are then selectively sampled and encoded to provide compact terrain information for low-level motion control. Given the history-conditioned state token s_t^{hist}, the elevation map M_t, and the future reference frames C_t^K, the encoder first employs a position selector, implemented as an MLP, to predict $N_g = 4$ glimpse locations:

$$\boldsymbol{r}_t^j = \left[\text{MLP}_\phi\left(\text{vec}(\boldsymbol{M}_t), \boldsymbol{s}_t^{\text{hist}}, \boldsymbol{C}_t^K\right)\right]_j, j \in \{1, \cdots, N_g\}, \tag{7}$$

where $\boldsymbol{r}_t^j \in \mathbb{R}^2$ denotes the j-th predicted sampling location and $\text{vec}(\cdot)$ denotes the vectorization operation that flattens its input into a one-dimensional vector. Then, we crop N_g terrain patches of size 5×5 centered at the predicted locations $\boldsymbol{r}_t^j$. Each patch, together with its center position $\boldsymbol{\rho}_t^j$ expressed in the robot frame, is passed through an MLP to obtain the local terrain token:

$$\boldsymbol{z}_t^j = \text{MLP}_\psi\left(\text{Crop}_{5\times5}(\boldsymbol{M}_t, \boldsymbol{r}_t^j), \boldsymbol{\rho}_t^j\right), j = \{1, \cdots, N_g\}. \tag{8}$$

Bilinear sampling makes the cropping operation differentiable, allowing the glimpse locations to be learned without region or foothold annotations. The resulting local terrain tokens are passed to the IFM.

Integration with IFM. In the first stage, the IFM produces the motion-intent token $\boldsymbol{s}_t^{\text{int}}$ conditioned solely on proprioceptive information and future motion references. At this stage, we inject terrain information into the IFM by augmenting the keys and values in Eq. (3) with the local terrain tokens $\{\boldsymbol{z}_t^j\}_{j=1}^{N_g}$: $K = V = \left[\boldsymbol{s}_t^{\text{hist}}; \boldsymbol{C}_t^K; \boldsymbol{z}_t^1; \ldots; \boldsymbol{z}_t^{N_g}\right]$. This enables the motion-intent token to incorporate terrain-aware information for low-level motion control.

Multi-Head Critic Extension. To encourage stable contacts, we introduce a new terrain-contact reward group r_t^{terrain} (summarized in Table I). Local height variation is used to evaluate touchdown quality, while contact labels obtained from offline terrain-mesh queries supervise the consistency between reference and simulated contacts. Additional penalties discourage foot slippage, stumbling, rapid contact switching, and excessive contact forces. To accommodate the new reward group r_t^{terrain}, we extend the multi-head critic from the first stage with an additional value head V_t^{terrain} that estimates the expected terrain-contact return. This is achieved by modifying the final layer of MLP_V to output four values instead of three. The values estimated by the multi-head critic in this stage are then given by

$$\boldsymbol{V}_t^{\text{inj}} = \left[V_t^{\text{upper}}, V_t^{\text{lower}}, V_t^{\text{terrain}}, V_t^{\text{aux}}\right]. \tag{9}$$

This dedicated value head provides a focused learning signal for terrain interaction and foothold selection, without interfering with the value estimates of the other reward groups.

2) Terrain-Adaptive Training Design: Beyond the architectural extension, the second stage requires several training-level adjustments so that the policy can deviate from the flat-ground reference where the terrain demands it.

Terrain-Aware Tracking Relaxation. Strictly penalizing every tracking error would suppress necessary changes in footholds, swing trajectories, and lower-body posture. We therefore introduce a terrain-aware relaxation into selected tracking objectives: $\widehat{e}_{t,h,j} = \left[e_{t,h,j} - \alpha_{h,j}\,\chi(\kappa_t)\,\tau_{m_h}(d_t)\right]_+$, where $e_{t,h,j}$ is the raw error of tracking objective h at link or joint j, $[\cdot]_+ = \max(0, \cdot)$, and $\widehat{e}_{t,h,j}$ replaces $e_{t,h,j}$ in the exponential tracking reward. Here, κ_t and d_t denote the terrain family and difficulty level of the tile under the robot: the indicator $\chi(\kappa_t) \in \{0, 1\}$ activates the relaxation only on slopes, stairs, and boxes; $\tau_{m_h}(d_t)$ is a base relaxation in the unit of objective h that grows linearly with d_t and saturates; and $\alpha_{h,j} \geq 0$ scales it per element, with $\alpha_{h,j} = 0$ recovering strict tracking. The relaxation is applied only to lower-body link position, link orientation, and joint position objectives, while upper-body tracking objectives and remaining motion constraints remain strict, preserving the reference motion while permitting necessary lower-body adaptation.

Global Position Correction. To prevent accumulated global position error, we feed the planar position error back into the reference velocity command. The error is computed in the reference anchor frame as $\boldsymbol{e}_{t,xy}^p = \left[(\boldsymbol{R}_t^{r,w})^\top(\boldsymbol{p}_t^{r,w} - \boldsymbol{p}_t^w)\right]_{xy}$, where $\boldsymbol{p}_t^{r,w}$ and $\boldsymbol{p}_t^w$ are the world-frame anchor positions of the reference and the robot, $\boldsymbol{R}_t^{r,w}$ is the world-frame rotation of the reference anchor, and $[\cdot]_{xy}$ extracts the planar components. The nominal planar reference velocity $\boldsymbol{v}_{t,xy}^r$ is then corrected as $\widetilde{\boldsymbol{v}}_{t,xy}^r = \boldsymbol{v}_{t,xy}^r + \text{clip}_{[-\bar{v},\bar{v}]}\left(g(\|\boldsymbol{v}_{t,xy}^r\|_2)\,\lambda_{\text{pos}}\,\boldsymbol{e}_{t,xy}^p\right)$, i.e., by a proportional term with gain λ_{pos}, modulated by a saturating speed gate $g(\cdot) \geq 0$ that grows with the nominal speed and vanishes for a stationary reference, and clipped component-wise at $\pm\bar{v}$. The corrected velocity serves as both the policy command and the velocity-tracking target. Setting $\lambda_{\text{pos}} = 0$ recovers pure velocity tracking, whereas $\lambda_{\text{pos}} > 0$ enables global position correction.

Terrain Curriculum and Reference Sampling. We train on flat terrain, slopes, stairs, boxes, and random rough terrain using a level-based curriculum, in which environments advance to harder terrain after successful completion and regress to easier levels after tracking failures. The terrain level jointly controls geometric difficulty, tracking relaxation, and selected termination delays. Since not every motion is feasible on every terrain, we use a coarse behavior–terrain compatibility rule to exclude clearly invalid combinations when sampling references. This requires neither terrain-adapted references, explicit foothold targets, nor paired motion–terrain demonstrations.

V. Experiments

We systematically evaluate PGMT in simulation and real-world deployment. In simulation, we compare it with general motion-tracking baselines and alternative terrain-perception designs across terrain types and difficulty levels. We then test it across multiple real-world control modes.

A. Experimental Setup

We evaluate terrain-adaptive motion tracking on five terrain families: flat terrain, slopes, stairs, box obstacles, and randomly rough terrain. Each family contains ten difficulty levels, denoted by L0–L9 in increasing order of difficulty. Each policy is evaluated over 9,600 matched 30-s episodes, with 192 episodes for each terrain–level pair. Corresponding episodes use identical motion references, starting frames, terrain assignments, dynamics randomization, observation corruption, and actuator-delay settings. An episode is considered successful if it reaches the maximum rollout horizon without early termination. Results are reported in Table II.

B. Comparison with Motion-Tracking Baselines

We compare PGMT with several general motion-tracking baselines, including **PGMT-Pretrain** before Perception Injection, our reimplementation **RGMT-Reimpl** [14], and the external checkpoint **SONIC v1.1 External** [16]. We also include the published Perceptive BFM result [3].

As shown in Table II, PGMT-Ours achieves completion rates of 87.81% over the full benchmark and 83.33% at L9, whereas PGMT-Pretrain reaches only 40.02% and 35.31%, respectively. The other motion trackers without terrain observations exhibit similar performance degradation. These results show that a general motion prior alone cannot reliably resolve mismatches between reference motions and local terrain, while terrain perception substantially improves tracking success on complex terrains.

TABLE II: **Evaluation performance and observed training cost.** Completion denotes reaching the 30-s horizon without early termination; Level 9 (L9) reports completion over 960 highest-difficulty episodes.

	Evaluation						Observed training cost		
Policy	Completion (%) ↑	L9 (%) ↑	Body Pos. (m) ↓	Body Ori. (°) ↓	Joint RMSE (rad) ↓	Contact F1 ↑	Time (s/iter.)	Max VRAM (GiB)	Envs/ GPU
Flat-only baselines on terrains									
PGMT-Pretrain	40.02	35.31	0.0897	27.70	0.2957	**0.7436**	4.54	16.11	10k
RGMT-Reimpl [14]	46.68	40.52	0.0700	20.15	**0.2251**	0.7680	5.42	24.78	8,192
SONIC v1.1 External† [16]	25.09	20.73	0.1204	36.42	0.2650	0.8232	–	–	–
Comparisons trained on terrains									
Perceptive BFM (paper-reported)* [3]	55.1	–	–	–	–	–	–	–	–
PGMT-NoHeight	86.53	78.85	0.0659	20.32	0.2937	0.8157	8.47	29.88	15k
PGMT-CNN	87.38	80.52	**0.0645**	18.33	0.2397	0.8173	11.95	27.30	15k
PGMT-2Glimpse	85.17	77.71	0.0676	18.27	0.2340	0.8234	8.36	28.10	15k
PGMT-Fixed	86.38	79.58	0.0645	18.47	0.2461	0.8218	8.28	28.66	15k
PGMT-Ours	**87.81**	**83.33**	0.0678	**18.14**	0.2299	0.8217	8.44	29.23	15k

RMSE denotes root-mean-square error. Training costs are reported only for runs trained in this work using four RTX 5090 GPUs; Time is the mean wall-clock time per PPO iteration, and VRAM is the peak single-GPU usage. †External-checkpoint results use re-audited adapters and checkpoint-native controllers. *Perceptive BFM results are quoted from the original work.

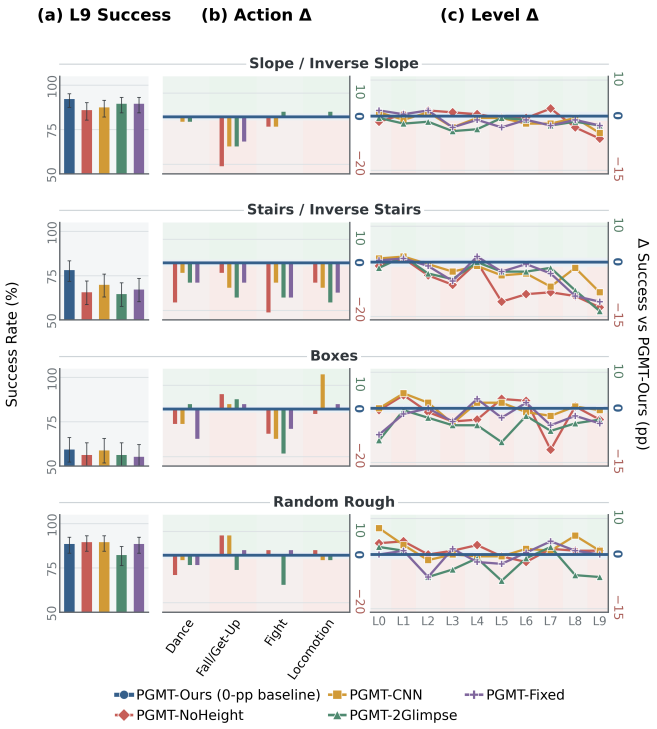

Fig. 3: **Terrain-perception ablations.** (a) L9 success with Wilson 95% intervals. (b) L9 motion-group success differences from PGMT-Ours (48 episodes per bar). (c) Success differences from PGMT-Ours across L0–L9.

C. Ablation Studies

We first evaluate four variants that modify only the terrain-perception mechanism. **PGMT-NoHeight** forces a zero input for terrain perception. **PGMT-CNN** replaces the glimpse module with a CNN that encodes the complete height scan into a fixed 2×2 terrain-token grid. **PGMT-2Glimpse** reduces the number of terrain glimpses from four to two. **PGMT-Fixed** uses four predefined glimpse locations instead of dynamically predicting them from the robot history and upcoming motion. All other policy components and training settings remain unchanged.

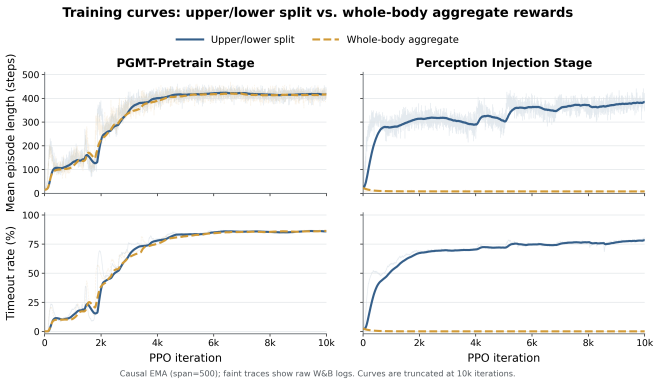

Fig. 4: **Split- and aggregate-return critics.** Both learn during Tracking Pretraining, but only the split-return critic progresses during Perception Injection. Curves show causal EMAs of the raw logs, truncated at 10k PPO iterations.

As shown in Table II, PGMT-Ours achieves the highest overall completion rate (87.81%) and L9 completion rate (83.33%). Across the four non-flat terrain families at L9 in Figure 3, PGMT-Ours achieves an average success rate of 79.56%, compared with 76.43% for the strongest alternative, PGMT-CNN. The largest advantage occurs on stairs, where PGMT-Ours achieves 78.12%, whereas the ablated variants achieve only 64.58–69.79%. Over the full evaluation schedule, PGMT-CNN remains comparable to PGMT-Ours. However, the larger separation on L9 stairs indicates that motion-conditioned spatial selection is particularly beneficial when complex terrain geometry and difficult motions occur together. These results further show that our adaptive region selection improve robustness on challenging terrains.

We further evaluate the effectiveness of the proposed multi-head split-return critic by replacing it with a single-head aggregate-return critic that estimates a single whole-

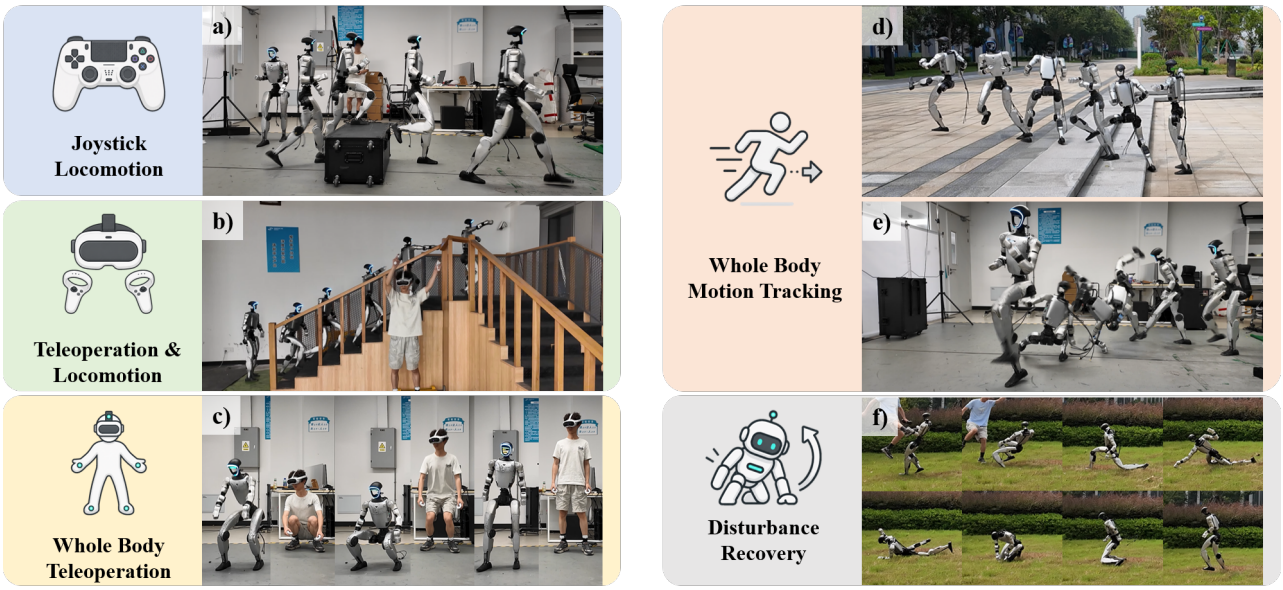

Fig. 5: **Unified real-world control with PGMT.** A single unified PGMT policy supports diverse command sources and control modes, including **(a)** joystick locomotion, **(b)** teleoperated locomotion, **(c)** whole-body teleoperation, **(d-e)** whole-body motion tracking, and **(f)** disturbance recovery. Locomotion commands are given through Unitree or PICO joystick. All behaviors are executed by the same policy without policy switching.

body return. The split-return critic uses a shared backbone and separate heads to estimate the returns associated with the upper-body, lower-body, and auxiliary reward groups. As shown in Figure 4, both critics successfully learn the terrain-agnostic motion-tracking policy during the Tracking Pretraining Stage, achieving comparable mean episode lengths and timeout rates. Once terrain perception is introduced, however, their training behaviors diverge substantially. The split-return critic continues to make sustained progress, eventually reaching a mean episode length of approximately 380 steps and a timeout rate of nearly 80%. By contrast, both metrics remain near zero for the aggregate-return critic.

D. Real-World Evaluation

We deploy the PGMT policy zero-shot on a 29-DoF Unitree G1. A Livox Mid-360S LiDAR and an onboard elevation-mapping pipeline provide terrain observations matching the simulation format. Perception and control run on a single NVIDIA Jetson Orin NX. Representative executions are shown in Fig. 1, and Fig. 5 summarizes the supported control modes.

1) Terrain-Adaptive Locomotion: We evaluate terrain-adaptive locomotion in both indoor and outdoor environments using flat-ground motion references generated online through motion matching. PGMT traverses uneven ground, ascends and descends staircases with unseen geometries, and climbs onto boxes up to $37\,\mathrm{cm}$ high. Across these terrains, the policy adapts foot placement, swing clearance, and body posture according to the local elevation map while following the locomotion intent.

We further evaluate the policy on irregular outdoor terrain, including grass and densely vegetated surfaces, as well as during jogging and sprinting on stairs, where the robot can clear multiple steps in a single stride (Fig. 1). Despite soft support surfaces and substantial perception noise, PGMT maintains stable locomotion without real-world fine-tuning.

2) Teleoperation: We demonstrate two teleoperation modes with the same PGMT policy: whole-body teleoperation and locomotion teleoperation. In whole-body teleoperation, the robot executes diverse operator-commanded behaviors, including, punching, squatting, lying down, and recovering to a standing posture. The operator's full-body motion is captured through PICO and retargeted online to the robot using GMR (Fig. 5c).

In locomotion teleoperation, the operator controls the robot's locomotion together with upper-body motion, while the lower-body reference is generated online through motion matching. As shown in Fig. 5b, the robot can climb a flight of stairs while the operator simultaneously teleoperates the upper body. PGMT adapts the locomotion reference to the local terrain, allowing the operator to control the upper body during terrain traversal without explicitly specifying terrain-dependent lower-body motions.

3) General Dynamic Whole-Body Motion Tracking: Beyond locomotion and teleoperation, we evaluate PGMT on diverse whole-body motions in the real world. PGMT retains the ability to execute highly dynamic motions (Fig. 5e) including cartwheels. We further evaluate terrain aware motion tracking non-flat terrain (Fig. 5d). PGMT executes these whole-body references on stairs and other uneven surfaces while adapting its contacts and posture to the local terrain. This demonstrates that terrain adaptation does not restrict PGMT to locomotion: the policy can preserve the structure of diverse whole-body motions while making the terrain-dependent adjustments required for stable execution.

4) Disturbance Robustness and Fall Recovery: Finally, we evaluate PGMT under large external disturbances (Fig. 5f). The robot withstands strong pushes and kicks while

performing a dance, and recovers its posture after substantial perturbations without interrupting the commanded motion.

When a sufficiently large disturbance causes a fall, PGMT autonomously recovers from the fallen configuration and returns to a stable standing posture. Both disturbance rejection and fall recovery are handled by the same unified policy without switching to a dedicated recovery controller, demonstrating robust whole-body control across both upright and fallen states.

VI. Conclusion

We presented PGMT, a perceptive general motion tracker that adapts terrain-agnostic reference motions to complex terrain without requiring terrain-matched demonstrations or explicitly generated reference trajectories. By injecting motion-conditioned terrain perception into a general tracking prior, PGMT extends general whole-body motion tracking beyond flat ground. Beyond general motion tracking, the same policy can serve directly as a terrain-adaptive locomotion controller or a whole-body teleoperation interface, and more broadly as a perceptive low-level controller that grounds terrain-agnostic motion commands from higher-level systems into physically feasible behaviors. A current limitation is that elevation maps encode geometry but not environmental semantics or affordances, making it difficult to distinguish obstacles from objects intended for interaction. Incorporating richer perceptual representations could further extend PGMT from terrain-aware motion execution toward more general humanoid interaction in unstructured environments.

References

[1] T. He, Z. Luo, W. Xiao, C. Zhang, K. Kitani, C. Liu, and G. Shi, "Learning human-to-humanoid real-time whole-body teleoperation," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 8944–8951.

[2] T. He, W. Xiao, T. Lin, Z. Luo, Z. Xu, Z. Jiang, J. Kautz, C. Liu, G. Shi, X. Wang *et al.*, "Hover: Versatile neural whole-body controller for humanoid robots," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 9989–9996.

[3] Z. Wang, Y. Li, T. Ma, Q. Zhang, Y. Fan, H. Xu, S. Yang, and J. Liang, "Perceptive behavior foundation model: Adapting human motion priors to robot-centric terrain," *arXiv preprint arXiv:2606.08059*, 2026.

[4] Q. Zhang, J. Ma, P. Liu, S. Shi, Z. Su, Z. Wang, J. Sun, W. Cui, J. Yu, G. Han *et al.*, "Meshmimic: Geometry-aware humanoid motion learning through 3d scene reconstruction," *arXiv preprint arXiv:2602.15733*, 2026.

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

[6] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "Amass: Archive of motion capture as surface shapes," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5442–5451.

[7] Z. Zhang, K. Wen, M. Xu, J. He, C. Li, T. Miki, C. Schwarke, C. Zhang, X. B. Peng, and M. Hutter, "Learning whole-body humanoid locomotion via motion generation and motion tracking," *IEEE Robotics and Automation Letters*, 2026.

[8] T. He, Z. Luo, X. He, W. Xiao, C. Zhang, W. Zhang, K. Kitani, C. Liu, and G. Shi, "Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning," *arXiv preprint arXiv:2406.08858*, 2024.

[9] M. Ji, X. Peng, F. Liu, J. Li, G. Yang, X. Cheng, and X. Wang, "Exbody2: Advanced expressive humanoid whole-body control, 2025," *arXiv preprint arXiv:2412.13196*.

[10] Z. Chen, M. Ji, X. Cheng, X. Peng, X. B. Peng, and X. Wang, "Gmt: General motion tracking for humanoid whole-body control," *arXiv preprint arXiv:2506.14770*, 2025.

[11] K. Yin, W. Zeng, K. Fan, M. Dai, Z. Wang, Q. Zhang, Z. Tian, J. Wang, J. Pang, and W. Zhang, "Unitracker: Learning universal whole-body motion tracker for humanoid robots," *IEEE Robotics and Automation Letters*, 2026.

[12] H. Zhao, S. Lin, Q. Ben, M. Dai, H. Fei, J. Wang, H. Zou, and J. Dong, "Smap: Self-supervised motion adaptation for physically plausible humanoid whole-body control," *arXiv preprint arXiv:2505.19463*, 2025.

[13] M. Yang, T. Yu, F. Li, and H. Chen, "Any2any: Efficient cross-embodiment transfer for humanoid whole-body tracking," *arXiv preprint arXiv:2605.23733*, 2026.

[14] Y. Ma, H. Yu, J. Xie, C. Lv, Q. Luo, C. Zhang, Y. Yin, B. Xing, X. Ren, and D. Zheng, "Robust and generalized humanoid motion tracking," *arXiv preprint arXiv:2601.23080*, 2026.

[15] T. He, J. Gao, W. Xiao, Y. Zhang, Z. Wang, J. Wang, Z. Luo, G. He, N. Sobanbab, C. Pan *et al.*, "Asap: Aligning simulation and real-world physics for learning agile humanoid whole-body skills," *arXiv preprint arXiv:2502.01143*, 2025.

[16] Z. Luo, Y. Yuan, T. Wang, C. Li, F. Castañeda, S. Chen, Z.-A. Cao, J. Li, D. Minor, Q. Ben *et al.*, "Sonic: Supersizing motion tracking for natural humanoid whole-body control," *Science Robotics*, vol. 11, no. 117, p. eaed4592, 2026.

[17] C. Tessler, Y. Guo, O. Nabati, G. Chechik, and X. B. Peng, "Maskedmimic: Unified physics-based character control through masked motion inpainting," *ACM Transactions On Graphics (TOG)*, vol. 43, no. 6, pp. 1–21, 2024.

[18] J. P. Araujo, Y. Ze, P. Xu, J. Wu, and C. K. Liu, "Retargeting matters: General motion retargeting for humanoid motion tracking," *arXiv preprint arXiv:2510.02252*, 2025.

[19] Z. Zhuang, S. Yao, and H. Zhao, "Humanoid parkour learning," *arXiv preprint arXiv:2406.10759*, 2024.

[20] W. Sun, B. Cao, L. Chen, Y. Su, Y. Liu, Z. Xie, and H. Liu, "Learning perceptive humanoid locomotion over challenging terrain," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025, pp. 6571–6578.

[21] J. Sun, G. Han, P. Sun, W. Zhao, J. Cao, J. Wang, Q. Zhang, and Y. Guo, "Dpl: Depth-only perceptive humanoid locomotion via realistic depth synthesis and cross-attention terrain reconstruction," *IEEE Robotics and Automation Letters*, 2026.

[22] S. Fu, Y. Zhang, Z. Cao, L. Yan, Y. Chen, Y. Yin, and Y. Gao, "Global-local attention decomposition for terrain encoding in humanoid perceptive locomotion," *arXiv preprint arXiv:2606.00637*, 2026.

[23] H. Wang, Z. Wang, J. Ren, Q. Ben, T. Huang, W. Zhang, and J. Pang, "Beamdojo: Learning agile humanoid locomotion on sparse footholds," *arXiv preprint arXiv:2502.10363*, 2025.

[24] P. Li, H. Li, M. Fan, F. Xu, S. Liao, Y. Ma, Z. Zeng, Z. Wang, Y. Jin *et al.*, "Taga: Terrain-aware active gaze learning for generalizable agile humanoid locomotion," *arXiv preprint arXiv:2606.05880*, 2026.

[25] P. Li, H. Li, Y. Ma, L. Chang, X. Yang, R. Yu, S. Liao, Y. Zhang, Y. Cao, Q. Zhu *et al.*, "Kivi: Kinesthetic-visuospatial integration for dynamic and safe egocentric legged locomotion," *arXiv preprint arXiv:2509.23650*, 2025.

[26] H. Han, C. Chen, L. Gong, X. Yang, H. Hu, J. Guo, Z. He, Y. Su, and F. He, "Guidewalk: Learning unified autonomous navigation and locomotion for humanoid robots across versatile terrains," *arXiv preprint arXiv:2606.10449*, 2026.

[27] Z. Wu, X. Huang, L. Yang, Y. Zhang, X. Chen, P. Abbeel, R. Duan, A. Kanazawa, C. Sferrazza, G. Shi *et al.*, "Perceptive humanoid parkour: Chaining dynamic human skills via motion matching," *arXiv preprint arXiv:2602.15827*, 2026.

[28] S. Chen, S. Zhao, Z. Wu, J. Li, G. Shi, and C. K. Liu, "Scenebot: Contact-prompted general humanoid whole body tracking with scene-interaction," *arXiv preprint arXiv:2606.27581*, 2026.

[29] Z. Ling, X. Yu, R. Yan, J. Cheng, Z. Wang, Q. Shuai, and C. Zou, "Gentrack: Physical alignment for robot-native motion generation and zero-shot humanoid tracking," *arXiv preprint arXiv:2608.01410*, 2026.

[30] 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 preprint arXiv:2508.08241*, 2025.

[31] F. G. Harvey, M. Yurick, D. Nowrouzezahrai, and C. Pal, "Robust motion in-betweening," *ACM Transactions on Graphics*, vol. 39, no. 4, pp. 60:1–60:12, 2020.