Moving Through Clutter: Scaling Data Collection and Benchmarking for 3D Scene-Aware Humanoid Locomotion via Virtual Reality

Beichen Wang, Yuanjie Lu, Linji Wang, Liuchuan Yu, and Xuesu Xiao

***Abstract*— Recent advances in humanoid locomotion have enabled dynamic behaviors such as dancing, martial arts, and parkour, yet these capabilities are predominantly demonstrated in open, flat, and obstacle-free settings. In contrast, real-world environments—homes, offices, and public spaces—are densely cluttered, three-dimensional, and geometrically constrained, requiring scene-aware whole-body coordination, precise balance control, and reasoning over spatial constraints imposed by furniture and household objects. However, humanoid locomotion in cluttered 3D environments remains underexplored, and no public dataset systematically couples full-body human locomotion with the scene geometry that shapes it. To address this gap, we present Moving Through Clutter (MTC), an open-source Virtual Reality (VR)-based data collection and evaluation framework for scene-aware humanoid locomotion in cluttered environments. Our system procedurally generates scenes with controllable clutter levels and captures embodiment-consistent, whole-body human motion through immersive VR navigation, which is then automatically retargeted to a humanoid robot model. We further introduce benchmarks that quantify environment clutter level and locomotion performance, including stability and collision safety. Using this framework, we compile a dataset of 348 trajectories across 145 diverse 3D cluttered scenes. The dataset provides a foundation for studying geometry-induced adaptation in humanoid locomotion and developing scene-aware planning and control methods.**

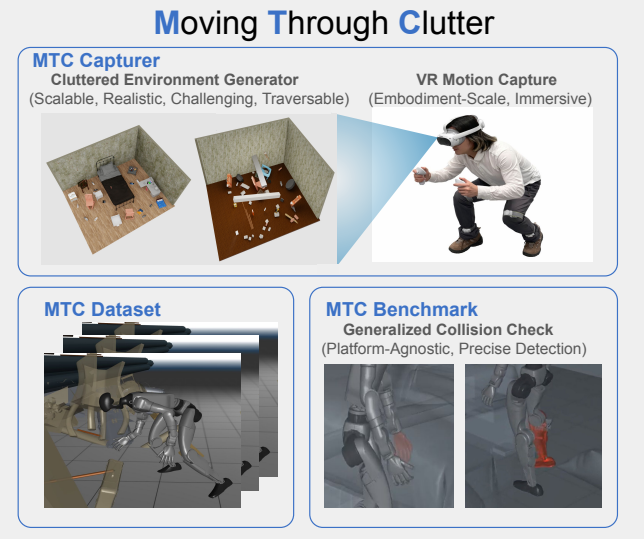

Fig. 1. We introduce MTC, a dataset and benchmark for humanoid locomotion in cluttered 3D environments. We first procedurally generate cluttered simulation scenes with diverse geometric constraints, then collect immersive, embodiment-scaled locomotion trajectories using VR-based full-body tracking within these environments. The resulting dataset enables benchmarking of scene-aware humanoid locomotion behaviors.

I. Introduction

Humanoid locomotion on flat terrain has witnessed rapid progress in recent years, with learning-based controllers achieving highly dynamic behaviors such as agile running, recovery from perturbations, and acrobatic motion synthesis [1], [2]. A central driver of this progress is the availability of large-scale human motion datasets. By leveraging motion priors derived from resources such as AMASS dataset [3], reinforcement learning policies can acquire diverse and natural locomotion strategies that would be difficult to engineer manually. The scale and diversity of motion data have emerged as a key factor shaping the capabilities of modern humanoid locomotion systems.

Despite these advances, most existing methods are developed and evaluated in open, obstacle-free settings. Real-world deployment, however, requires humanoid robots to operate within cluttered environments where locomotion is constrained by surrounding three-dimensional (3D) geometry, like in homes or offices. Locomotion in such settings demands continuous whole-body postural adaptation coupled to environmental structure, including lateral clearance management, height-aware motion, and acyclic and asymmetric limb adjustment to avoid collisions. These requirements fundamentally differ from flat-ground locomotion and remain largely underexplored from a data-centric perspective.

Although human motion datasets have proven essential for humanoid locomotion learning, no publicly available resource currently provides scene-aware, embodiment-consistent humanoid locomotion trajectories that are explicitly aligned with 3D cluttered environments. The primary bottleneck lies in scalable data acquisition. Traditional motion capture pipelines rely on open studios without occlusions and therefore fail to encode interactions with spatial constraints. Constructing diverse physical environments, e.g., with large furniture and household objects, for data collection is expensive and difficult to reproduce at scale. Meanwhile, recent Virtual Reality (VR)-based teleoperation systems demonstrate that immersive interfaces can bridge human motion and robot embodiment [4], [5], yet these efforts are primarily designed for real-time control rather than systematic dataset construction.

Motivated by these limitations, we introduce an open-

All authors are with George Mason University, 4400 University Dr, Fairfax, VA 22030, USA {bwang25, ylu22, lwang44, lyu20, xiao}@gmu.edu

This work has taken place in the RobotiXX Laboratory at George Mason University. RobotiXX research is supported by National Science Foundation (NSF, 2350352), Army Research Office (ARO, W911NF2320004, W911NF2420027, W911NF2520011), Air Force Research Laboratory (AFRL), US Air Forces Central (AFCENT), Google DeepMind (GDM), Clearpath Robotics, Raytheon Technologies (RTX), Tangenta, Mason Innovation Exchange (MIX), and Walmart.

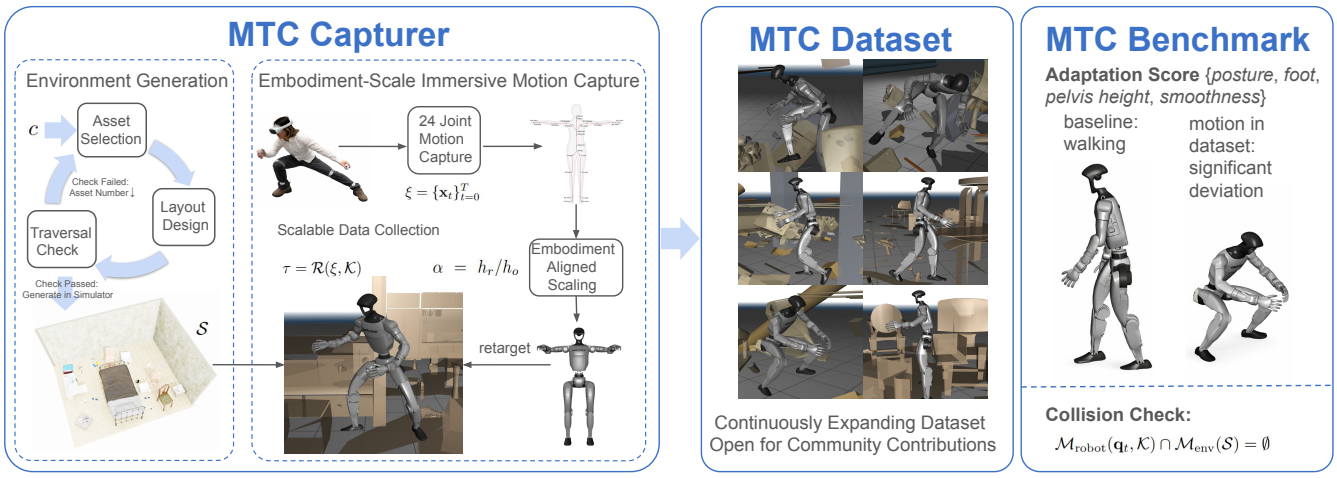

Fig. 2. System overview of **MTC**. **MTC Capturer**: immersive VR-based system for collecting embodiment-scaled human locomotion trajectories in cluttered environments. **MTC Dataset**: a large-scale collection of locomotion trajectories recorded in procedurally generated cluttered scenes. **MTC Benchmark**: evaluation framework for measuring scene-aware locomotion behaviors relative to normal walking.

source framework, Moving Through Clutter (MTC), for collecting and evaluating scene-aware humanoid locomotion data. Our approach combines procedural cluttered environment generation with embodiment-scaled virtual reality capture. By scaling the human operator to match the physical proportions of the target humanoid during data acquisition, the recorded trajectories become geometrically consistent with robot-scale locomotion, effectively transforming human-scale demonstrations into embodiment-aligned motion data. The entire pipeline operates in virtual environments without requiring physical scene construction, enabling scalable and reproducible dataset generation under controlled clutter level and geometric difficulty. To systematically validate the collected trajectories, we further introduce a quantitative evaluation benchmark that measures locomotion difficulty and collision safety under geometric constraints, providing a standardized protocol for assessing scene-aware humanoid locomotion. The contributions of our MTC framework are summarized as follows:

- MTC Capturer: An embodiment-scaled VR capture paradigm that produces geometrically consistent locomotion trajectories aligned with humanoid embodiment and procedurally generated 3D scenes with continuous clutteredness parameterization;
- MTC Dataset: An open-source dataset that jointly provides whole-body locomotion trajectories and corresponding 3D scene configurations; and
- MTC Benchmark: A dual evaluation benchmark that quantitatively measures locomotion difficulty and collision safety in geometrically constrained environments.

II. Related Works

We review related work in learning humanoid locomotion, human motion datasets and VR teleoperation, and procedural simulation environment generation.

A. Learning-Based Humanoid Locomotion

Reinforcement learning has enabled significant progress in humanoid locomotion. Radosavovic et al. [6] and Haarnoja et al. [7] demonstrated robust sim-to-real transfer and agile dynamic behaviors, establishing that learning-based controllers can achieve stable and versatile locomotion. Human motion priors further accelerate this progress: DeepMimic [8] and Adversarial Motion Priors (AMP) [9] introduced example-guided and style-based rewards derived from human demonstrations, while ExBody [10], ExBody2 [11], and SONIC [5] scaled whole-body tracking with large motion datasets.

Despite these advances, existing policies are predominantly trained and evaluated on flat or sparsely structured terrain, where the primary challenge is balance and contact stability. Even terrain-variant settings typically consist of engineered primitives such as slopes or isolated obstacles. In contrast, locomotion in cluttered 3D indoor environments requires continuous modulation of torso orientation, lateral clearance, and full-body posture to avoid multi-height collisions. Current approaches achieve dynamic locomotion yet rarely condition control on detailed scene geometry, leaving geometry-conditioned traversal strategies largely unexplored.

While Xue et al. [12] make an important step toward cluttered indoor traversal by introducing HumanoidPF as an informative representation and training traversal behaviors purely with RL, their formulation is primarily optimized for collision avoidance learnable via strong, hand-designed directional guidance—HumanoidPF is explicitly queried at body parts as policy observation and also shapes dense reward signals. As a result, the learned behaviors are naturally biased toward a compact set of goal-directed traversal skills (e.g., hurdle/crouch/squeeze) that can be efficiently acquired under such guidance, rather than toward rich, human-like, and reusable geometry-conditioned locomotion strategies grounded in large-scale demonstrations. This leaves open how to systematically capture and leverage diverse full-body locomotion data paired with the exact 3D scene geometry

that induces it, so that policies can be learned not only to be collision-free, but also to be expressive, natural, and broadly reusable across clutter distributions.

B. Human Motion Datasets and VR Teleoperation

Large-scale motion capture datasets such as AMASS [3], BABEL [13], HumanML3D [14], and Motion-X [15] provide diverse human motion sequences and semantic annotations, substantially advancing motion modeling. However, these datasets are recorded in obstacle-free studios and do not encode explicit coupling between motion and environmental geometry, preventing models from learning how locomotion adapts to spatial constraints.

Scene-aware datasets including PROX [16], SAMP [17], HUMANISE [18], and TRUMANS [19] introduce motion aligned with 3D indoor scenes. Yet they focus primarily on human motion prediction and generation characterized by quasi-static interactions from the computer vision perspective, limiting applicability to humanoid locomotion with different kinematics and actuation constraints, particularly for sustained traversal through narrow or irregular passages.

VR-based teleoperation systems [4], [20]–[24] and retargeting frameworks [25], [26] demonstrate effective human-to-humanoid motion transfer. However, these efforts emphasize real-time control for loco-manipulation rather than systematically constructing reusable locomotion datasets paired with procedurally varied environments and quantitative benchmarking protocols.

C. Procedural Simulation Environment Generation

Procedural environment generation has primarily supported embodied navigation and visual understanding. ProcTHOR [27], Infinigen [28], [29], Holodeck [30], and Habitat [31] generate large-scale interactive indoor or outdoor scenes with semantic coherence and navigational accessibility.

However, humanoid locomotion in cluttered indoor spaces is constrained by full-body volumetric clearance, multi-height collision risks, and kinematic feasibility. Existing procedural generators neither model embodiment-aware traversal constraints nor parameterize scene difficulty in terms of clearance margins or collision risk. As a result, they are not designed to support geometry-conditioned humanoid locomotion learning or benchmarking.

Taken together, prior work has advanced locomotion control, motion modeling, teleoperation, and scene generation largely in isolation. What remains missing is a unified framework that couples procedural clutter generation, immersive embodiment-scaled locomotion data capture scalable by VR, motion retargeting, and quantitative benchmarking for scene-aware humanoid traversal in cluttered 3D environments.

III. MTC Capturer

We now formalize the 3D scene-aware humanoid locomotion task and describe the MTC Capturer, the core data acquisition module that enables embodiment-scaled motion capture within generated cluttered environments in VR.

A. Problem Formulation

We consider goal-directed humanoid locomotion in cluttered 3D environments. Let a humanoid robot with kinematic model $\mathcal{K}$ operate in a scene $\mathcal{S}$. The robot configuration space is denoted by C, where a configuration $\mathbf{q} \in C$ encodes all joint configurations of the robot. A goal-directed humanoid locomotion task is specified by an initial configuration $\mathbf{q}_0 \in C$ and a goal set $\mathcal{G} \subset C$. A trajectory is defined as a time-ordered sequence of configurations

$$\tau = \{\mathbf{q}_t\}_{t=0}^{T}, \qquad \mathbf{q}_t \in C,$$

which is considered successful if its terminal configuration satisfies

$$\mathbf{q}_T \in \mathcal{G}.$$

A traversal trajectory is considered valid only if it satisfies the following necessary conditions throughout execution:

- **Collision safety.** The robot must remain collision-free with respect to the scene geometry at all times:

$$\mathcal{M}_{\text{robot}}(\mathbf{q}_t, \mathcal{K}) \cap \mathcal{M}_{\text{env}}(\mathcal{S}) = \emptyset, \quad \forall t \in \{0, \dots, T\}.$$

  Here $\mathcal{M}_{\text{robot}}(\mathbf{q}_t, \mathcal{K})$ denotes the robot body mesh obtained via forward kinematics and $\mathcal{M}_{\text{env}}(\mathcal{S})$ denotes the environment geometry.
- **Postural stability.** The robot must maintain dynamic support throughout execution and must not undergo irrecoverable loss of balance.

We denote by $\mathcal{T}(\mathcal{S}, \mathcal{K}, \mathcal{G})$ the set of trajectories that satisfy both goal completion defined by $\mathcal{G}$ and the above physical feasibility conditions defined by $\mathcal{S}$ and $\mathcal{K}$.

B. Data-Driven Solutions with MTC Capturer

Learning-based approaches require a humanoid robot dataset of embodiment-specific traversal trajectories,

$$\mathcal{D}^{\text{robot}} = \{(\mathcal{S}_i, \mathcal{K}, \mathcal{G}_i, \tau_i)\}_{i=1}^{N}, \quad \tau_i \in \mathcal{T}(\mathcal{S}_i, \mathcal{K}, \mathcal{G}_i),$$

where each τ_i is a physically feasible, goal-directed humanoid locomotion trajectory executed under the kinematic embodiment $\mathcal{K}$ in a cluttered scene $\mathcal{S}_i$.

However, constructing $\mathcal{D}^{\text{robot}}$ directly with real humanoid platforms is impractical. First, constructing geometry-conditioned locomotion data requires a diverse distribution of scene geometries $\mathcal{S}$ that impose structured 3D spatial constraints. Physical construction of such environments is costly and difficult to regulate systematically. Second, collecting collision-free and stable humanoid trajectories in such environments presupposes solving the scene-aware locomotion problem under study; moreover, human demonstration from direct teleoperation of highly articulated humanoids under tight geometric constraints is extremely challenging.

To overcome these limitations, we introduce the **MTC Capturer**, a virtual-reality data collection pipeline that enables scalable scene generation and embodiment-consistent motion acquisition in VR. The MTC Capturer first procedurally generates geometrically diverse cluttered scenes in simulation to enable scalable environment construction,

avoiding the need to set up any physical geometry in the real world. After that, it captures embodiment-scaled human reference motions as operators traverse these virtual environments, allowing geometric constraints to induce whole-body adaptation at the target robot's proportions. Let

$$\xi = \{\mathbf{x}_t\}_{t=0}^{T}$$

denote a human reference motion sequence collected. The resulting motion is subsequently mapped to the humanoid configuration space through a retargeting function

$$\tau = \mathcal{R}(\xi, \mathcal{K}),$$

yielding robot traversal trajectories consistent with embodiment $\mathcal{K}$. MTC Capturer's objective is therefore to construct a reproducible embodiment-consistent dataset that couples procedurally generated scene geometries with immersive human reference motions,

$$\mathcal{D}^{\mathrm{MTC}} = \{(\mathcal{S}_i, \mathcal{K}, \mathcal{G}_i, \xi_i)\}_{i=1}^{N},$$

where each sample consists of a generated scene $\mathcal{S}_i$ and an embodiment-scaled human reference motion ξ_i. Corresponding robot trajectories are obtained via retargeting.

The two components of MTC Capturer are detailed in the following subsections. The first component addresses scalable scene construction, while the second enforces embodiment-consistent geometry during motion acquisition.

C. Procedural Environment Generation

A locomotion-relevant cluttered scene $\mathcal{S}$ must satisfy three essential requirements:

- It must preserve the semantic characteristics of a real-world scenario. Scenes should reflect recognizable room types (e.g., bedroom, living room, and kitchen) with functionally meaningful object arrangements rather than arbitrary object scattering;
- It must exhibit sufficient geometric clutter such that traversal cannot be accomplished by trivial straight-line walking. Instead, the environment should induce constraint-driven whole-body adaptation, such as side-stepping through narrow passages, torso reorientation, crouching under height-constrained obstacles, or clearance-aware stepping; and
- It must guarantee at least one feasible traversal path from a start to a goal for the target humanoid.

MTC Capturer addresses these three requirements through a structured procedural pipeline described below.

1) Geometric Regimes: Each scene is generated under one of two geometric regimes. The first structured domestic regime models semantically organized indoor layouts dominated by furniture-induced lateral confinement and corridor-like free space. The second debris-style regime introduces irregular obstacle configurations that create both planar entanglement and vertical clearance restrictions. In addition to dense and non-axis-aligned ground-level clutter, this regime incorporates overhead beams and structural elements that constrain height clearance, jointly inducing behaviors such as crouching, ducking, or crawling. Representative examples of the two regimes are shown in Fig. 3.

Fig. 3. Examples of cluttered environments from two geometric regimes in MTC: structured domestic layouts (left) and debris-style layouts (right).

2) Semantic Layout Structure: Each scene begins by sampling a geometric regime and room type. To preserve structural coherence while enabling controllable variation, assets are organized into functional tiers and placed in a hierarchical order:

- **Anchor Element Layer:** A dominant structural or functional component that defines the primary spatial organization of the scene. In structured domestic layouts, this corresponds to a functional core object (e.g., bed, sofa, and dining table), around which the room is organized. In the debris regime, the anchor role is fulfilled by major structural elements (e.g., load-bearing pillars or primary beam assemblies) that determine the global obstruction pattern. The anchor element is instantiated first according to regime-specific placement priors, ensuring plausible spatial structure.
- **Supporting Large Elements Layer:** Large objects or structures that establish the dominant topology of the environment. In domestic layouts, these include wall-aligned furniture such as wardrobes, cabinets, or bookshelves. In the debris regime, these may include secondary pillars or extended structural members. Placement is performed via probabilistic sampling conditioned on room type and size tier, followed by rejection checks to avoid overlap.
- **Small Clutter Layer:** Density-controlling objects that perturb navigable free space at a finer scale. These include freestanding items (e.g., chairs) as well as scattered small obstacles. Small clutter is injected after the primary layout has been established, allowing density to vary without altering global structure.
- **Vertical Obstruction Layer (debris regime only):** Additional structural members (e.g., overhead beams, rebars, and I-shaped steel) are attached to pillars or walls to introduce height-clearance constraints. These elements create coupled planar and vertical restrictions, inducing crouching, ducking, or crawling behaviors beyond lateral navigation.

During generation, anchor elements are instantiated first, followed by supporting large elements, and finally small clutter and regime-specific obstructions. This hierarchical layout strategy ensures that global spatial structure is determined before local density perturbations are introduced, preserving semantic plausibility while progressively shaping

the navigable topology.

Two diagonally opposite spawn zones, denoted as $\mathcal{Z}_{\text{start}}$ and $\mathcal{Z}_{\text{goal}}$, are reserved as obstacle-free regions prior to placement, defining an origin–destination traversal pair.

3) Clutterness-Driven Density Control: To provide continuous and interpretable control over scene density, we introduce a scalar parameter $c \in [0, 1]$ representing the target floor-occupancy ratio. Rather than manually specifying discrete object counts, the number of items injected into each placement layer is computed from the room area and the typical footprint size of assets in that layer.

Let A denote the room floor area. We allocate the clutterness budget only to non-anchor placement layers. Specifically, for each density-controlled layer $i \in \{\texttt{Supporting Large Elements}, \texttt{Small Clutter}\}$, the target number of items is determined by

$$n_i = \left\lfloor \frac{c\, A\, w_i}{\bar{a}_i} \right\rfloor, \quad \sum_i w_i = 1.$$

Here, w_i is the weight that distributes the occupancy budget between large structural elements and small clutter, while the anchor element is instantiated independently of w_i and c. $\bar{a}_i$ denotes the mean ground-plane footprint area of all assets that can be assigned for that layer. The footprint $\bar{a}_i$ is computed at runtime from each asset's axis-aligned bounding box on the floor plane. This formulation ensures that the number of items scales proportionally with room size, and that larger assets naturally result in fewer instances under the same clutterness parameter c due to their increased footprint.

While the clutterness parameter c controls obstacle density in expectation, high-density configurations may inadvertently eliminate feasible traversal paths for the target morphology. We therefore perform an explicit navigability verification step after object placement.

4) Morphology-Aware Navigability Verification: After object placement, scene feasibility is validated through a 2D grid-based reachability test. The floor is discretized at a fixed resolution. Before rasterization, each obstacle footprint is inflated by a configurable humanoid clearance radius K, derived from the embodiment descriptor $\mathcal{K}$, effectively constructing a morphology-aware configuration-space map. The occupancy grid is then constructed from these inflated obstacles, and Breadth First Search is performed on the resulting free-space map. The search is initialized from all free cells within $\mathcal{Z}_{\text{start}}$, and the scene is considered traversable if any cell within $\mathcal{Z}_{\text{goal}}$ can be reached. If no feasible traversal path exists, the scene is not discarded outright. Instead, we invoke a constraint-preserving annealed resampling procedure to restore connectivity while maintaining structural coherence.

5) Constraint-Preserving Annealed Resampling: If a candidate layout fails the reachability check, the system applies an annealing schedule that progressively removes assets while preserving structural elements. Small clutter—the most expendable layer—is reduced first via multiplicative decay (20% per annealing level), as these elements primarily perturb local free space without defining global structure. If connectivity remains unsatisfied, supporting large elements are subsequently reduced starting from the next annealing level using the same decay rate. Regime-defining anchor elements and major structural furniture are preserved until higher annealing stages, ensuring that the global spatial organization of the scene remains intact as long as the scene is physically navigable. This hierarchical relaxation policy defines a monotonic density schedule and guarantees convergence to a feasible layout given an appropriate scene size within a bounded number of retries while preserving structural coherence.

Following annealed resampling, the realized clutterness level may differ from the originally specified target c. We therefore define the realized density

$$c' = \frac{\sum_k a_k}{A},$$

where a_k is the ground-plane footprint of each non-anchor object in the final layout. While c controls density in expectation during sampling, c' reflects the actual floor occupancy ratio after reachability-constrained annealing.

Collectively, hierarchical layout construction, morphology-aware verification, and constraint-preserving annealing yield semantically structured, density-controlled environments whose navigability is explicitly conditioned on the target humanoid embodiment.

D. Embodiment-Scaled Immersive Motion Capture

Standard motion-capture datasets record human motion at natural body proportions. When retargeted to a robot with different stature, geometric clearances experienced during capture may no longer correspond to those encountered by the robot, leading to unintended collisions or infeasible postures. To mitigate this embodiment mismatch, we enforce embodiment-consistent geometry during data collection.

To match the scale of the humanoid and human operator, let h_r denote the standing height of the target robot and h_o the measured standing height of the human operator. We define the embodiment scale factor as $\alpha = h_r/h_o$. During data collection, the virtual environment is rendered with a uniform scale factor $1/\alpha$, such that spatial clearances experienced by the operator correspond to those encountered by the robot at its embodiment scale. Joint poses are recorded at human scale and uniformly rescaled by the factor α after capture, yielding an embodiment-consistent motion sequence expressed in the robot's spatial scale.

To match the skeletal structure of the humanoid and human operator, existing retargeting framework can be employed. For this work, we use General Motion Retargeting framework of Ze et al. [4] without modification.

IV. MTC Dataset and Benchmark

We introduce the MTC Dataset and Benchmark to support learning approaches and systematic evaluation of scene-aware humanoid locomotion in cluttered environments.

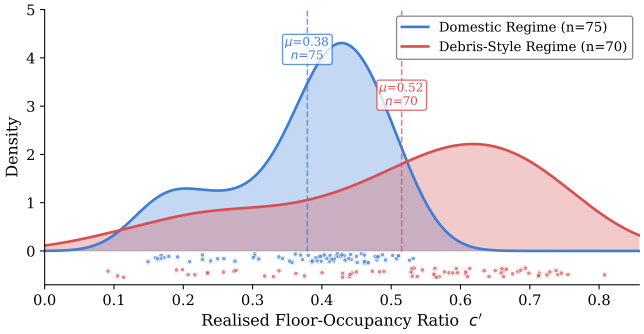

Fig. 4. Distribution of realised floor-occupancy ratio c' across 145 generated scenes. Individual scene values are shown as jittered strips below the density curves; dashed lines indicate per-regime means.

A. MTC Dataset

1) Dataset Overview: The MTC dataset consists of 145 procedurally generated scenes $\{\mathcal{S}_k\}$ and 348 associated traversal trajectories $\{\xi_j\}$ collected under the Unitree G1 humanoid embodiment. Scenes span both geometric regimes, i.e., domestic and debris-style, and multiple semantic room types, e.g., bedroom, living room, and kitchen. While the target clutterness parameter c is defined over $[0, 1]$, empirical analysis under the G1 embodiment indicates that realised densities c' in the range $[0.2,\ 0.6]$ most frequently correspond to geometrically challenging yet traversable layouts.

Human motion is captured using a PICO 4 Ultra VR system with integrated full-body tracking, providing 24-joint skeletal pose measurements per frame. In total, the current dataset contains approximately 731,000 motion frames across 348 trajectories, corresponding to roughly 2.3 hours of humanoid locomotion data. Individual trajectories range from 431 to 6,939 frames in length, with a mean of 2,101 frames. The MTC dataset is actively expanding and the MTC capturer pipeline will be open-sourced to encourage community contributions.

2) Scene Density Distribution: To characterize geometric variability across the dataset, we analyze the distribution of realised clutterness levels c' over all generated scenes. Fig. 4 shows the kernel density estimates of c' for both regimes, computed as the fraction of room floor area occupied by non-anchor object footprints on a 5 cm rasterization grid.

3) Case Study: Goal-Conditioned Route Diversity: To illustrate the richness of geometry-induced behavior, we analyze a representative scene under four goal configurations. Fig. 5 presents the scene floor plan together with the ground-plane projections of pelvis trajectories, along with representative snapshots of the corresponding obstacle-avoidance behaviors. Markers $\bullet$, $\star$, and $\times$ denote start locations, goal positions, and obstacle avoidance maneuvers, respectively.

Although all motions occur within the same environment, different goal placements induce distinct traversal routes that expose the agent to different local geometric constraints—such as narrow passages, low-clearance structures, or densely cluttered regions. This example illustrates that behavioral diversity in MTC arises not only from varying scene layouts, but also from goal-conditioned routing within a single environment.

B. MTC Benchmark

1) Motion Adaptation Score: Locomotion under geometric constraints induces coordinated deviations from nominal flat-ground walking across multiple kinematic dimensions. To quantify geometry-induced whole-body adaptation, we evaluate trajectory statistics within four complementary subspaces:

- *Posture*: pelvis-relative joint positions and velocities, capturing configuration-level whole-body adjustments;
- *Vertical motion*: pelvis height, vertical velocity, and vertical acceleration, reflecting height-clearance adaptation;
- *Foot interaction*: foot heights and vertical velocities, characterizing stepping and obstacle negotiation; and
- *Smoothness*: third-order positional differences (jerk) across joints, measuring dynamic modulation under constraint.

For each subspace, frame-wise feature vectors are extracted and summarized by their empirical mean and covariance. A reference distribution is estimated from a corpus of temporally aligned, speed-normalized flat-ground walking trajectories and approximated by a multivariate Gaussian. Given a test trajectory, we compute the Fréchet distance between its empirical feature distribution and the baseline distribution within the same subspace. The Fréchet distance captures both mean displacement and covariance shift, enabling joint evaluation of first- and second-order kinematic deviations.

Formally, let (μ_r, Σ_r) and (μ_t, Σ_t) denote the reference and test feature statistics, respectively. The squared Fréchet distance is computed as

$$d^2 = \|\mu_r - \mu_t\|^2 + \mathrm{Tr}\Big(\Sigma_r + \Sigma_t - 2(\Sigma_r^{1/2}\Sigma_t\Sigma_r^{1/2})^{1/2}\Big).$$

Subspace distances are normalized and aggregated via uniform weighting to produce a scalar adaptation score. Higher values indicate greater deviation from nominal flat-ground locomotion. The score is intended as a relative deviation measure rather than an absolute proxy for task difficulty.

2) Collision Safety Assessment: Physical feasibility within cluttered 3D geometry is treated as an explicit evaluation criterion in the MTC Benchmark. Given an evaluated trajectory $\{\mathbf{q}_t\}_{t=1}^{T}$ and scene mesh $\mathcal{M}$, collision safety is assessed via signed distance queries against the original (non-convex) scene geometry.

For each frame t, forward kinematics maps the joint configuration $\mathbf{q}_t$ to world-frame link poses. Surface sample points are generated on each link and queried against a precomputed signed distance field of $\mathcal{M}$. Let $s_{t,i}$ denote the signed distance of sample point i at frame t, where negative values indicate penetration into the scene geometry. The frame-wise penetration depth is then defined as

$$d_t = \max\big(0,\ -\min_i s_{t,i}\big),$$

so that $d_t = 0$ indicates a collision-free configuration and $d_t > 0$ measures the maximum instantaneous penetration

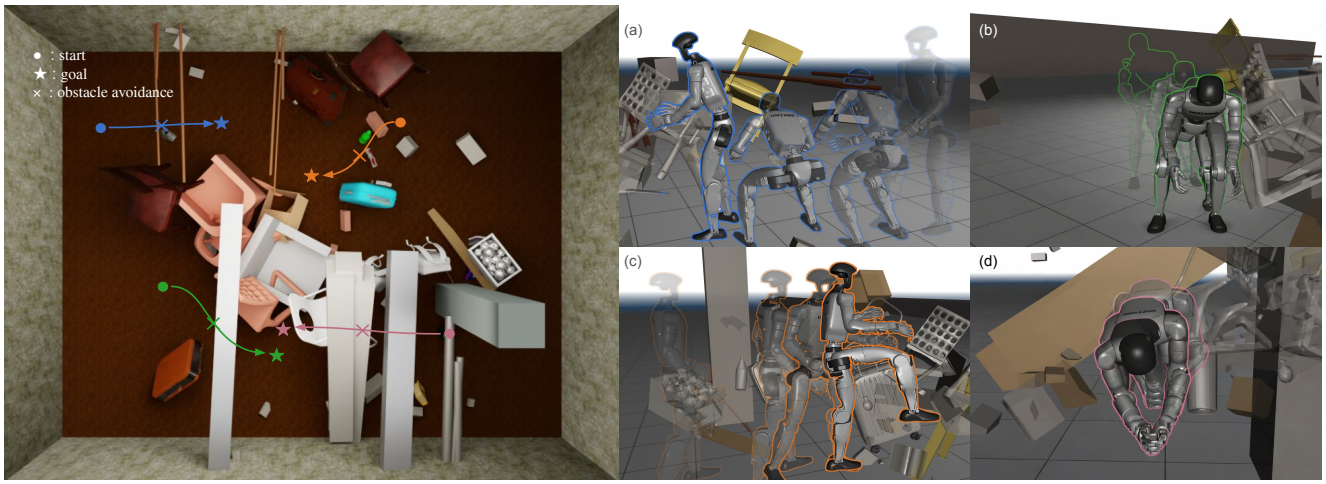

Fig. 5. Case study of goal-conditioned route diversity in a representative MTC scene. The left panel visualizes the scene floorplan together with the ground-plane projections of pelvis trajectories under four goal configurations. Markers ●, ⋆, and × denote start locations, goal positions, and obstacle-avoidance maneuvers, respectively. Insets (a–d) show representative locomotion behaviors observed along the corresponding routes: (a) crouched lateral shuffling, (b) crouched forward shuffling, (c) high-knee lateral step-over, and (d) prone crawling.

depth at frame t. Based on $\{d_t\}_{t=1}^{T}$, the benchmark reports four quantitative safety metrics:

$$R_{\text{col}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}[d_t > 0],$$
$$d_{\max} = \max_t d_t,$$
$$\bar{d}_{\text{cond}} = \frac{\sum_t d_t \mathbf{1}[d_t > 0]}{\max(1, \sum_t \mathbf{1}[d_t > 0])},$$
$$I_{\text{pd}} = \frac{1}{T}\sum_{t=1}^{T} d_t.$$

Here, R_{col} measures collision frequency over the trajectory, $d_{\max}$ captures the worst-case geometric violation, $\bar{d}_{\text{cond}}$ quantifies average penetration severity conditioned on collision events (defined as zero when no collision occurs), and I_{pd} reflects time-normalized penetration magnitude across the full traversal. Together, these metrics characterize both discrete collision occurrence and continuous penetration severity under full scene geometry. The evaluation code will be released as part of the MTC framework to ensure reproducibility.

3) Dataset-Level Benchmark Statistics: To analyze the behavioral coverage of the collected trajectories, we examine the distribution of per-frame kinematic features across four subspaces, defined by MTC benchmark. For each subspace, feature vectors are extracted from all trajectories as well as from a baseline recording of unobstructed level-ground walking, and projected onto the first two principal components for visualization (Fig. 6).

In the *posture* subspace, baseline frames form a compact closed orbit consistent with a regular gait cycle. In the *foot* subspace, the baseline distribution concentrates into a narrow, phase-structured wedge with two symmetric lobes, reflecting alternating left–right foot interactions rather than a closed loop. Similarly, in the *pelvis height* subspace, baseline frames concentrate within a narrow band corresponding to the nearly constant pelvis height of normal walking. Across these subspaces, the dataset distributions extend substantially beyond the baseline patterns, indicating diverse posture and foot-interaction adaptations induced by cluttered environments.

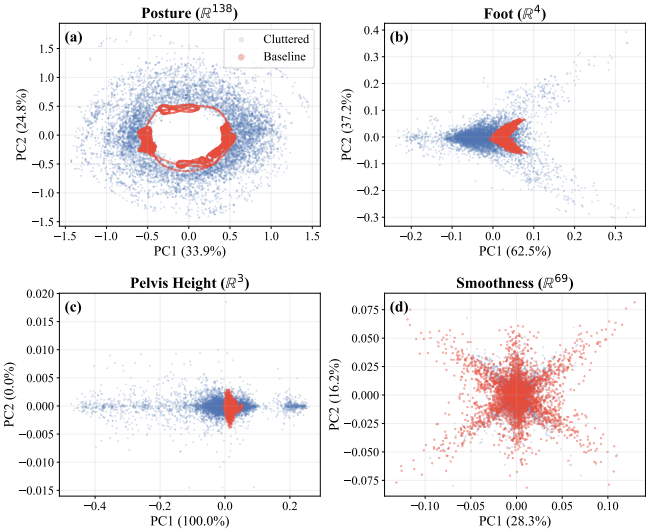

Fig. 6. PCA projections of per-frame kinematic features for four subspaces. Red: baseline level-ground walking; Blue: cluttered-environment trajectories.

In the *smoothness* subspace, the baseline and dataset distributions largely overlap, indicating that the collected trajectories remain consistently smooth despite geometric constraints—a property not always achieved by learned robot policies.

Together, these projections confirm that the dataset captures a broad spectrum of posture and foot-interaction adaptations while maintaining high motion quality, providing a challenging and diverse benchmark for locomotion in cluttered environments.

V. Conclusion and Limitations

We presented MTC, a framework that comprises a data capturer, a dataset, and a benchmark for data-driven, scene-

aware humanoid locomotion in cluttered 3D environments. The MTC Capturer leverages immersive VR to capture human locomotion in virtual 3D cluttered environments at scale. The MTC Dataset couples procedurally generated geometric regimes with embodiment-consistent traversal trajectories, while the MTC Benchmark evaluates trajectories along two complementary axes: geometry-induced kinematic adaptation and collision safety under full scene geometry.

Through dataset-level analysis and goal-conditioned case studies, we demonstrate that MTC captures diverse whole-body adaptation induced by rich spatial constraints, rather than merely diversifying different scene layouts. Quantitative results show that distinct geometric regimes elicit measurably different traversal behaviors from human demonstrators, revealing structured adaptation patterns that existing benchmarks, which vary only scene layout or goal placement, are insufficient to characterize. Furthermore, the proposed benchmark metrics provide a principled protocol for analyzing locomotion performance under spatial constraints, offering actionable guidance for downstream algorithm development.

Preliminary results show that a reinforcement learning-based motion tracking policy, trained to imitate MTC trajectories, can reproduce geometry-induced traversal behaviors with low collision rates.

Despite these contributions, several limitations remain. First, the current pipeline employs scene-agnostic retargeting; fully scene-aware motion generation remains an open challenge requiring advances in control and learning. Second, scene layout generation relies on manually designed placement priors rather than learned distributions, which may not fully capture real-world variability. Integrating generative scene models such as Vision-Language Models could improve realism and diversity. Third, the dataset focuses on locomotion-centric traversal and does not model contact-assisted progression, which may be necessary in extreme clutter where multi-contact support is required for balance. Finally, the VR-based motion capture relies on pose estimation and may introduce tracking noise. Our future work will incorporate high-precision motion capture to further improve accuracy.

In summary, MTC provides a foundation for systematic study of geometry-induced adaptation in humanoid locomotion and stimulates further research on scene-aware humanoid planning and control with scalable data collection.

REFERENCES

[1] D. Kim, J. Bae, J. Lee, D. Son, and S. Oh, "Stage-wise reward shaping for acrobatic robots: A constrained multi-objective reinforcement learning approach," *arXiv preprint*, 2024.

[2] Y. Wang, S. Zhu, P. Zhi, Y. Li, J. Li, Y.-L. Li, Y. Xiao, X. Wang, B. Jia, and S. Huang, "OmniXtreme: Breaking the generality barrier in high-dynamic humanoid control," *arXiv preprint*, 2026.

[3] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "AMASS: Archive of motion capture as surface shapes," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 5442–5451.

[4] Y. Ze, S. Zhao, W. Wang, A. Kanazawa, R. Duan, P. Abbeel, G. Shi, J. Wu, and C. K. Liu, "TWIST2: Scalable, portable, and holistic humanoid data collection system," *arXiv preprint*, 2025.

[5] Z. Luo, Y. Yuan, T. Wang, C. Li, S. Chen, F. Castaneda, Z.-A. Cao, J. Li *et al.*, "SONIC: Supersizing motion tracking for natural humanoid whole-body control," *arXiv preprint*, 2025.

[6] I. Radosavovic, T. Xiao, B. Zhang, T. Darrell, J. Malik, and K. Sreenath, "Real-world humanoid locomotion with reinforcement learning," *Sci. Robot.*, vol. 9, no. 89, p. eadi9579, 2024.

[7] T. Haarnoja, B. Moran, G. Lever, S. H. Huang, D. Tirumala, J. Humplik, M. Wulfmeier, S. Tunyasuvunakool, N. Y. Siegel, R. Hafner *et al.*, "Learning agile soccer skills for a bipedal robot with deep reinforcement learning," *Sci. Robot.*, vol. 9, no. 89, p. eadi8022, 2024.

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

[9] X. B. Peng, Z. Ma, P. Abbeel, S. Levine, and A. Kanazawa, "AMP: Adversarial motion priors for stylized physics-based character animation," *ACM Trans. Graph.*, vol. 40, no. 4, pp. 1–20, 2021.

[10] X. Cheng, Y. Ji, J. Chen, R. Yang, G. Yang, and X. Wang, "Expressive whole-body control for humanoid robots," in *Proc. Robot., Sci. Syst. (RSS)*, 2024.

[11] M. Ji, X. Peng, F. Liu, J. Li, G. Yang, X. Cheng, and X. Wang, "ExBody2: Advanced expressive humanoid whole-body control," *arXiv preprint*, 2024.

[12] 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," *arXiv preprint*, 2025.

[13] A. R. Punnakkal, A. Chandrasekaran, N. Athanasiou, A. Quiros-Ramirez, and M. J. Black, "BABEL: Bodies, action and behavior with English labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 722–731.

[14] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng, "Generating diverse and natural 3D human motions from text," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 5152–5161.

[15] J. Lin, A. Zeng, S. Lu, Y. Cai, R. Zhang, H. Wang, and L. Zhang, "Motion-X: A large-scale 3D expressive whole-body human motion dataset," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023.

[16] M. Hassan, V. Choutas, D. Tzionas, and M. J. Black, "Resolving 3D human pose ambiguities with 3D scene constraints," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 2282–2292.

[17] M. Hassan, D. Ceylan, R. Villegas, J. Saito, J. Yang, Y. Zhou, and M. J. Black, "Stochastic scene-aware motion prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 11 374–11 384.

[18] Z. Wang, Y. Chen, T. Liu, Y. Zhu, W. Liang, and S. Huang, "HUMANISE: Language-conditioned human motion generation in 3D scenes," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022.

[19] N. Jiang, Z. Zhang, H. Li, X. Ma, Z. Wang, Y. Chen, T. Liu, Y. Zhu, and S. Huang, "TRUMANS: Scaling up dynamic human-scene interaction modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

[20] 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 *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2024.

[21] 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," in *Proc. Conf. Robot Learn. (CoRL)*, 2024.

[22] X. Cheng, J. Li, S. Yang, G. Yang, and X. Wang, "Open-TeleVision: Teleoperation with immersive active visual feedback," in *Proc. Conf. Robot Learn. (CoRL)*, 2024.

[23] Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn, "HumanPlus: Humanoid shadowing and imitation from humans," in *Proc. Conf. Robot Learn. (CoRL)*, 2024.

[24] Y. Ze, Z. Chen, J. P. Araujo, Z.-a. Cao, X. B. Peng, J. Wu, and C. K. Liu, "TWIST: Teleoperated whole-body imitation system," *arXiv preprint*, 2025.

[25] J. P. Araujo, Y. Ze, P. Xu, J. Wu, and C. K. Liu, "Retargeting matters: General motion retargeting for humanoid motion tracking," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2026.

[26] L. Yang, X. Huang, Z. Wu, A. Kanazawa, P. Abbeel, C. Sferrazza, C. K. Liu, R. Duan, and G. Shi, "OmniRetarget: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction," *arXiv preprint*, 2025.

[27] M. Deitke, E. VanderBilt, A. Herrasti, L. Weihs, J. Salvador, K. Ehsani, W. Han, E. Kolve, A. Farhadi, A. Kembhavi, and R. Mottaghi,

"ProcTHOR: Large-scale embodied AI using procedural generation," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022.

[28] A. Raistrick, L. Lipson, Z. Ma, L. Mei, M. Wang, Y. Zuo, K. Kayan, H. Wen, B. Han, Y. Wang, A. Newell, H. Law, A. Goyal, K. Yang, and J. Deng, "Infinite photorealistic worlds using procedural generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 12 630–12 641.

[29] A. Raistrick, L. Mei, K. Kayan, D. Yan, Y. Zuo, B. Han, H. Wen, M. Parakh, S. Alexandropoulos, L. Lipson, Z. Ma, and J. Deng, "Infinigen indoors: Photorealistic indoor scenes using procedural generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

[30] Y. Yang, F.-Y. Fan, K. Dickinson, J. Wu, D. Khashabi, Y. Choi, and A. Kembhavi, "Holodeck: Language guided generation of 3D embodied AI environments," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

[31] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, D. Parikh, and D. Batra, "Habitat: A platform for embodied AI research," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 9339–9347.