Visual Imitation Enables Contextual Humanoid Control

Author Names Omitted for Anonymous Review.

Fig. 1: **VIDEOMIMIC** is a real-to-sim-to-real pipeline that converts monocular videos into transferable humanoid skills, letting robots learn context-aware behaviors (terrain-traversing, stairs-climbing, sitting) in a single policy. Video results are available on our webpage: https://videomimic.github.io/.

***Abstract*—How can we teach humanoids to climb staircases and sit on chairs using the surrounding environment context? Arguably, the simplest way is to *just show them*—casually capture a human motion video and feed it to humanoids. We introduce VIDEOMIMIC, a real-to-sim-to-real pipeline that mines everyday videos, jointly reconstructs the humans and the environment, and produces whole-body control policies for humanoid robots that perform the corresponding skills. We demonstrate the results of our pipeline on real humanoid robots, showing robust, repeatable contextual control such as staircase ascents and descents, sitting and standing from chairs and benches, as well as other dynamic whole-body skills—all from a single policy, conditioned on the environment and global root commands. VIDEOMIMIC offers a scalable path towards teaching humanoids to operate in diverse real-world environments.**

I. Introduction

How do we learn to interact with the world around us—like sitting on a chair or climbing a staircase? We watch others perform these actions, try them ourselves, and gradually build up the skill. Over time, we can handle new chairs and staircases, even if we have not seen those exact ones before. If humanoid robots could learn in this way—by observing everyday videos—they could acquire diverse contextual whole-body skills without relying on hand-tuned rewards or motion-capture data for each new behavior and environment. We refer to this ability to execute environment-appropriate actions as contextual control.

We introduce VideoMimic, a real-to-sim-to-real pipeline that turns monocular videos—such as casual smartphone captures—into transferable skills for humanoids. From these videos, we jointly recover the 4D human-scene geometry, retarget the motion to a humanoid, and train an RL policy to track the reference trajectories. We then distill the policy into a single unified policy that observes only proprioception, a local height-map, and the desired root direction. This distilled policy outputs low-level motor actions conditioned on the terrain and body state, allowing it to execute appropriate behaviors—such as stepping, climbing, or sitting—across unseen environments without explicit task labels or skill selection.

We develop a perception module that reconstructs 3D human motion from a monocular RGB video, along with aligned scene point clouds in the world coordinate frame. We convert the point clouds into meshes and align them with gravity to ensure compatibility with physics simulators. The global motion and local poses are retargeted to a humanoid with constraints that ensure physical plausibility, accounting for the embodiment gap. The mesh and retargeted data seed a goal-conditioned DeepMimic [32]-style reinforcement-learning phase in simulation: we warm-start on MoCap data, then train a single policy to track motions from multiple videos in their respective height-mapped environments while randomizing mass, friction, latency, and sensor noise for robustness. Once our tracking policy is trained, we distill it using DAgger [41] to a policy that operates without conditioning on target joint angles. The new policy observes proprioception, an 11×11 height-map patch centered on the torso, and the vector to the goal in the robot's local reference frame. PPO fine-tuning under this reduced observation set yields a generalist controller that, given height-map and root direction at test time, selects and smoothly executes context-appropriate actions such as stepping, climbing, or sitting. In particular, every step of our policy relies only on observations available at real-world deployment, making it immediately runnable on real hardware.

Our approach bridges 4D video reconstruction and robot skill learning in a single, data-driven loop. Unlike earlier work that recovers only the person or the scene in isolation, we jointly reconstruct both at a physically meaningful scale and represent them as meshes and motion trajectories suitable for physics-based policy learning. We train our approach on 123 monocular RGB videos, which will be released. We validate the approach through deployment on a real Unitree G1 robot, which shows generalized humanoid motor skills in the context of surrounding environments, even on unseen environments. We will release the reconstruction code, policy training framework, and the video dataset to facilitate future research.

II. Related Work

Method	Env. Real-to-Sim	Context. Ctrl	Real Robot
DeepMimic / SfV [32, 33]	✗	✗	✗
Egocentric Loco [1]	✗	✓	✓
ASAP [7]	✗	✗	✓
Humanoid Loco. [34]	✗	✗	✓
H2O / ExBody2 [6, 13]	✗	✗	✓
Parkour [8, 63]	✗	✓	✓
VideoMimic (Ours)	✓	✓	✓

TABLE I: **Comparison of methods across different features.** VideoMimic transfers both human motion and scene geometry from real videos to simulation, learns context-aware control in simulation, and successfully deploys the resulting policy on real-world environments.

Learning Skills on Legged Robots. Recent progress in legged-robot motor skills follows two complementary streams. Reward-based methods use model-free RL in simulation, shaping behavior with handcrafted objectives that mix task terms (e.g., velocity tracking) and motion-naturalness regularizers; thanks to massive parallel physics engines [11, 27], this paradigm has produced agile locomotion on quadrupeds and humanoids without motion data. However, each new behavior demands tuning of user-defined rewards and environment scripting [12, 18, 17, 1, 34, 8, 49, 22]. Data-driven methods instead imitate reference motion, originally MoCap clips or monocular video, training a simulated character to track them and porting the idea to robots [32, 33, 56, 25, 6, 13]. For example, recent work [35, 36] frames legged locomotion as a next- token prediction task and pre-trains a policy on human data in kinematic space, showing strong performance. While imitation bypasses reward engineering [7], existing works typically assume flat ground or manually designed setups, limiting context-aware whole-body control; even animation systems that model human-scene contact rely on instrumented MoCap stages and thus lack scalability [5, 30]. Our system conditions on visual observations, the local height-maps, and learns environment-aware skills such as stair-climbing and chair-sitting directly from monocular RGB videos. Joint 4D human–scene reconstruction provides physically consistent reference motions, which RL distills into policies that transfer to a real humanoid (Table I).

Human and Scene Reconstruction from Images and Videos. Early monocular-video methods regress pose and shape of humans [23] in a camera-relative frame with deep networks [23, 14, 15], which suffices for rendering, action recognition, or single-person tracking [31, 28, 37, 26, 38] but leaves the global trajectory—and thus context-aware dynamics—undefined; pioneers like SfV hand-tuned a global

scale and even assumed a static camera, limiting generality. Recent methods combine human motion priors with SfM/SLAM to recover metric trajectories [54, 58, 16], yet still model only the person and camera. Advances in general scene parsing [20, 61, 50] have enabled joint human-scene reconstruction that resolves scale via multi-view cues or learned priors [29, 21], but these systems have not been validated on robots. Parallel work injects physics constraints in post-processing or simulation [57, 59, 47, 62, 19], trading scalability for realism. Our pipeline unifies these threads: it simultaneously estimates metric human motion and surrounding geometry from in-the-wild videos—without MoCap, pre-scanned scenes, or reward engineering—and outputs simulator-ready trajectories that respect contacts and collisions, enabling scalable learning of whole-body humanoid skills.

III. Real-to-Sim Data Acquisition

Our real-to-sim pipeline proceeds as summarized in Figure 2. We extract per-frame human poses and a raw scene point cloud from the input video (Sec. III-A); jointly optimize them to obtain metrically aligned human trajectories and scene geometry (Sec. III-B); apply gravity alignment and convert the filtered point cloud into a lightweight mesh (Sec. III-C); and retarget the refined trajectories to the humanoid under joint-limit, contact, and collision constraints. The resulting motion–mesh pairs are ready for policy learning in Sec. IV.

A. Preprocessing

We preprocess monocular RGB videos with off-the-shelf state-of-the-art human pose estimation and SfM methods. First, people are detected and associated across frames using Grounded SAM2 [39, 40]. For each detected person, we recover per-frame 3D SMPL [23] parameters with VIMO [51], obtaining per-frame local pose θ^t, shape β, and SMPL 3D joints $J^t_{\text{3D}} \in \mathbb{R}^{J\times 3}$. We detect 2D keypoints J^t_{2D}, i.e., body joint pixel positions, with ViTPose [53]. Foot contact is regressed by BSTRO [9]. For scene reconstruction, we obtain the world point cloud from either MegaSaM [20] or MonST3R [61], which is parameterized as per-frame depth D^t, camera pose $[R^t|t^t]$, and a shared camera intrinsic matrix K. Note that the resulting point cloud is not metrically accurate.

To coarsely position the person in the world frame, we follow the initialization strategy of SLAHMR [54], using (i) the camera focal length predicted by SfM and (ii) the ratio between the average 2D limb length from the ViTPose detections $\tilde{J}^t_{\text{2D}}$ and the corresponding metric scale 3D limb length in J^t_{3D}, we estimate a similarity factor per frame that yields a coarse global trajectory $(\phi^{t_0}, \gamma^{t_0})$. Separately, we also lift $\tilde{J}^t_{\text{2D}}$ to 3D by un-projecting each pixel (u, v) with its depth $D^t_{u,v}$ and intrinsics K from SfM: $\tilde{J}^t_{\text{3D}} = K^{-1}[u, v, 1]^\top D^t_{u,v}$. The lifted joints are then used to jointly optimize human poses and scene geometry scale, as described in the following section.

B. Joint Human–Scene Reconstruction

Our pipeline jointly optimizes the human trajectory and the scene scale. The variables are the humans' global translations $\gamma^{1:T}$, global orientations $\phi^{1:T}$, local poses $\theta^{1:T}$, and the scene point-cloud scale α. Because MegaSam pointclouds are scale-ambiguous, the metric human height prior in the SMPL body models serves as the metric reference, while the lifted joints $\tilde{J}^t_{\text{3D}}$ refine both the global trajectory and the local pose. We therefore solve for α simultaneously, reconciling any residual mismatch between the human-derived scale and the scene geometry.

Inspired by He et al. [6], we optionally run a scale-adaptation pass that searches for an SMPL shape $\beta^\star$ whose height and limb proportions match those of the G1 robot, prior to joint human-scene optimization. The SMPL joints are then extracted from the reshaped mesh. This use of a prefitted G1-scale SMPL β effectively rescales the scene geometry to G1 size, improving the feasibility of humanoid motion—e.g., enabling actions like running or climbing over large obstacles—and facilitating reference motion learning. For real-world deployment, we skip this step and operate directly on the original metric-scale scene.

The objective combines joint-distance losses in 3D (L_{3D}), computed as the L1 distance between $\tilde{J}^t_{\text{3D}}$ and J^t_{3D}, and 2D projection losses (L_{2D}), along with a temporal smoothness regularizer (L_{Smooth}) that discourages frame-to-frame jitter:

$$\arg\min_{\alpha,\gamma,\phi,\theta} \; w_{\text{3D}} L_{\text{3D}} + w_{\text{2D}} L_{\text{2D}} + L_{\text{Smooth}}.$$

We optimize this objective with a Levenberg–Marquardt solver implemented in JAX [55]. Running on an NVIDIA A100 GPU, the optimizer processes a 300-frame sequence in approximately 20 ms after compilation.

C. Generating Simulation-Ready Data

To deploy the monocular reconstruction in a physics engine, we (i) align it with real-world gravity using GeoCalib [48] and (ii) convert the noisy, dense point cloud into a lightweight mesh that imposes meaningful geometric constraints and supports memory-efficient parallel training. We use NKSR [10] for meshification.

IV. Policy Learning

Given the kinematic reference from our clips and scenes, our policy learning pipeline produces a context-conditioned policy that can perform skills from the references when prompted by the appropriate environmental context. Figure 3 gives an overview of our pipeline, detailed below.

Policy Learning. We use Proximal Policy Optimization [43] implementation from Rudin et al. [42] for training our policy. Our learning takes place in the IsaacGym simulator [27].

Observations. Our policies are conditioned on both proprioceptive and target-related observations. The proprioceptive inputs include a history of the robot's joint positions (q), joint velocities ($\dot{q}$), angular velocity (ω), projected gravity vector (g), and previous actions (a^{t-1}); we use a history

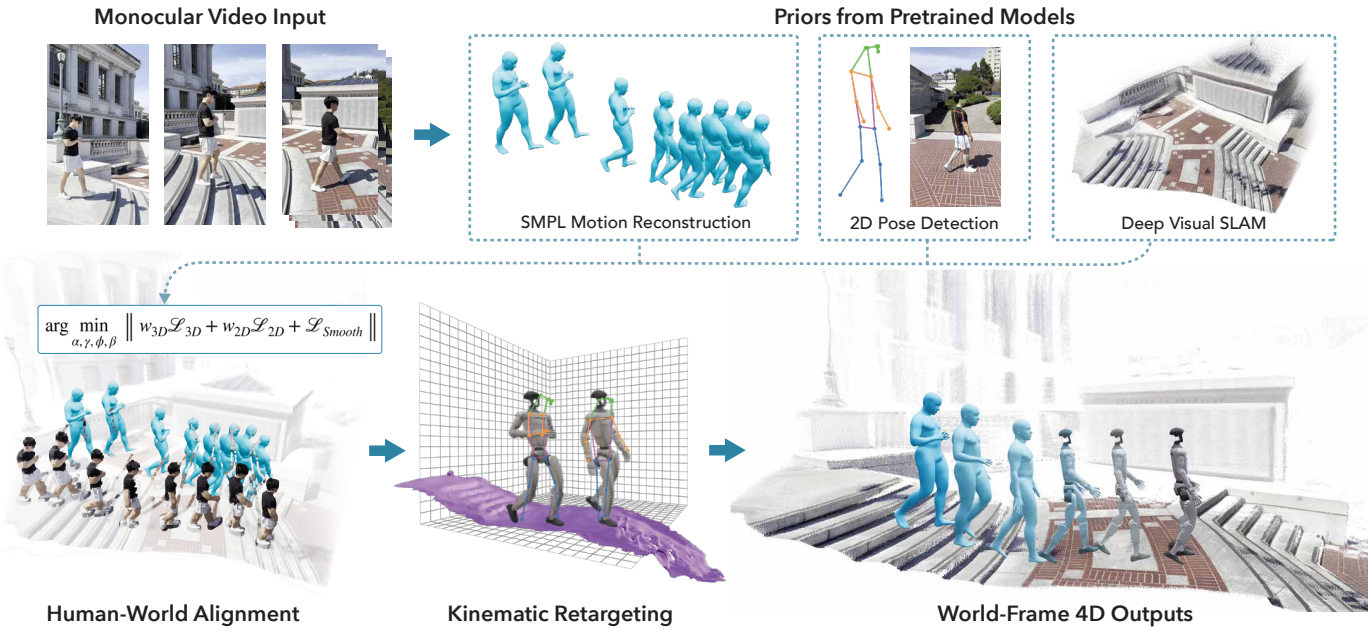

Fig. 2: **VideoMimic Real-to-Sim.** A casually captured phone video provides the only input. We first reconstruct per-frame human motion and 2D keypoints, along with a dense scene point cloud. An efficient optimization jointly aligns the motion and point cloud, recovers statistically accurate metric scale using a human height prior, and registers the human trajectory based on human-associated points. The point cloud is then converted to a mesh, aligned with gravity, and the motion is retargeted to a humanoid in the reconstructed scene. This yields world-frame trajectories and simulator-ready meshes that serve as inputs for policy training.

length of 5 in practice. In addition, the policy receives local target observations: the target joint angles, target root roll and pitch, and the desired root direction, specified by relative x-y offset and yaw angle between the robot's current root position and the target root, all expressed in the robot's local frame. For policies conditioned on heightmaps, we further provide an elevation map around the torso. This is represented as an 11×11 grid sampled at 0.1m intervals, which captures local terrain geometry. Finally, the critic receives additional privileged observations.

Batched Tracking. Our system utilizes a batched variant of DeepMimic [32] in order to learn to imitate motions using RL. We implement Reference State Initialization [32] in addition to motion load balancing similar to Tessler et al. [46], upweighting motions with a lower success rate.

Rewards. Our RL reward is designed entirely around data-driven tracking terms—specifically, link and joint positions, joint velocities, and foot-contact signals—so that raw demonstrations can be translated into physically executable motions with minimal hand-tuning. We have two objectives: (1) reducing reliance on manually crafted priors that are typically introduced through reward engineering, and (2) ensuring physical feasibility of the resulting motions. These two goals can conflict: because the reference trajectories are purely kinematic data from humans, exact tracking may result in non-physical motion. We therefore introduce an action-rate penalty along with several other penalty criteria designed to discourage exploiting simulator physics. We train our policy over the stages described below.

Stage 1: MoCap Pre-Training. MoCap pre-training lets a policy learn challenging skills from noisy video reconstruction while keeping hand-crafted priors to a minimum and bridging the human-to-robot embodiment gap. Earlier work tackles this either by sampling start poses with multi-agent RL [33] or by having a privileged simulator imitate the motion [7]. Radosavovic et al. [36] and Singh et al. [45] instead employed a form of kinematic pre-training on human data. We adopt a simpler yet effective strategy: first pre-train the policy on MoCap trajectories, then fine-tune it on our reconstructed video data—both stages use reinforcement learning in a physics simulator. Even the MoCap-only policy can be deployed directly on the real robot. We used LAFAN motion capture data [4] retargeted to Unitree G1. For the pretrained policies, the conditioning the policy receives is the target joint angles, target root roll/pitch, and desired root direction.

Stage 2: Scene-Conditioned Tracking. After MoCap pre-training, we initialize the policy from MPT checkpoints and introduce scene awareness by conditioning on the environment heightmap. The heightmap is integrated via a projection into the MPT policy's latent space residually with an initial weight of 0. We then randomly sample motions and perform DeepMimic-style tracking across reconstructed terrains. During this stage, the policy continues to receive motion-specific tracking conditioning, including target joint angles, root roll/pitch, and desired root directions.

Stage 3: Distillation. Following the stage of batched tracking, we distill via DAgger [41] to a policy that does not observe target joint angles or root roll/pitch observations. We are then able to use the desired root directions observations as conditioning signals to control the robot's position, which can

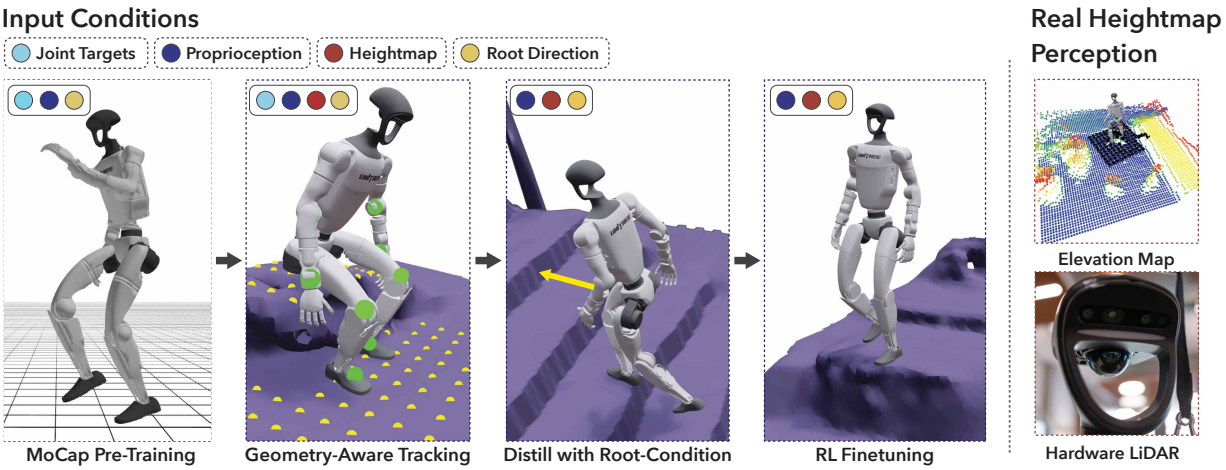

Fig. 3: **Policy training in sim.** Our pipeline of training RL starts with a dataset of Motion Capture trajectories. We then inject a heightmap observation and track whole-body reference trajectories from our videos in various environments. We proceed to distill a policy conditioned only on the root position of the robot. We then finetune this policy directly with RL using the same reduced observation set. Our pipeline is motivated by three goals: (a) producing motions that are fast and faithful to the original video demonstrations; (b) ensuring observations are available in real-world settings; and (c) training a generalist policy that distills knowledge from all video demonstrations into a single model applicable beyond the training set.

be fed either from a joystick or potentially a path provided by a high-level controller. In this way, our framework unifies the previously separate approaches of joystick tracking and global reference following. Our distilled policy benefits from the fact that the teacher is also trained with observation randomization, hence it learns actions under some uncertainty, which would not be the case if we started by training with full body observations; this has been shown to be helpful in other contexts with policy learning [24].

Stage 4: Under-conditioned RL Finetuning. After distilling our policies to be exclusively conditioned on the root of our trajectories, we perform another round of RL. This is because behaviours which are learned conditioned on target joints may be sub-optimal for policies which are not conditioned on such targets. In practice, we found that this can significantly boost performance as compared to distilled policies. It also makes it possible to add lower-quality reference motions to the reference set since removing targets from the actor in effect makes it a "data-driven" reward signal with an under-constrained actor which is able to follow references appropriate to context.

V. Results

We demonstrate that humanoid robots can learn context-aware skills in diverse environments by imitating everyday human videos. We first evaluate the robustness of our reconstruction pipeline against baselines. Next, we demonstrate its versatility, highlighting its potential impact on future research. We then detail our curated video dataset. Finally, we ablate the MPT component and present demonstrations successfully transferred from simulation policies to a physical robot.

A. Reconstruction and Data

Evaluation. We evaluate the robustness of our reconstruction pipeline on a subset of the SLOPER4D dataset [2], assessing both human trajectory and scene geometry reconstruction.

Methods	WA-MPJPE	W-MPJPE	Chamfer Distance
WHAM* [44]	189.29	1148.49	–
TRAM [51]	149.48	954.90	10.66
Ours	**112.13**	**696.62**	**0.75**

TABLE II: **Comparison of Reconstruction.** *: WHAM does not recover the environment.

We compare our method against baselines following the standard evaluation protocol [44, 51]. As summarized in Table II, our method consistently achieves the best performance, outperforming prior work in both human trajectory accuracy (WA/W-MPJPE) and scene geometry (Chamfer Distance).

Versatility. Figure 4 highlights the breadth of our reconstruction pipeline, showcasing (i) robust environment reconstruction from an Internet video involving dynamic human-scene interaction, (ii) multi-human reconstruction and retargeting. Furthermore, the dense point cloud reconstruction enables ego-view RGB-D rendering via simple rasterization. While not used in our current policy, this offers a promising direction for future work—especially given the challenges of rendering naturalistic images in simulation.

Video Data. We curated 123 casually recorded smartphone videos of people performing everyday activities in diverse indoor and outdoor settings, including sitting, standing up from furniture, walking up/down stairs (even backwards), and stepping onto blocks.

MPT ablation. We ablate the impact of pretraining on motion capture data. MPT has multiple effects: first, reference motions are noisy and thus harder to learn to track tabula rasa. Second, initial positions of the robot are often not entirely statically stable or may have some interpenetrations with the scene. Hence, MPT can help stabilize learning during the

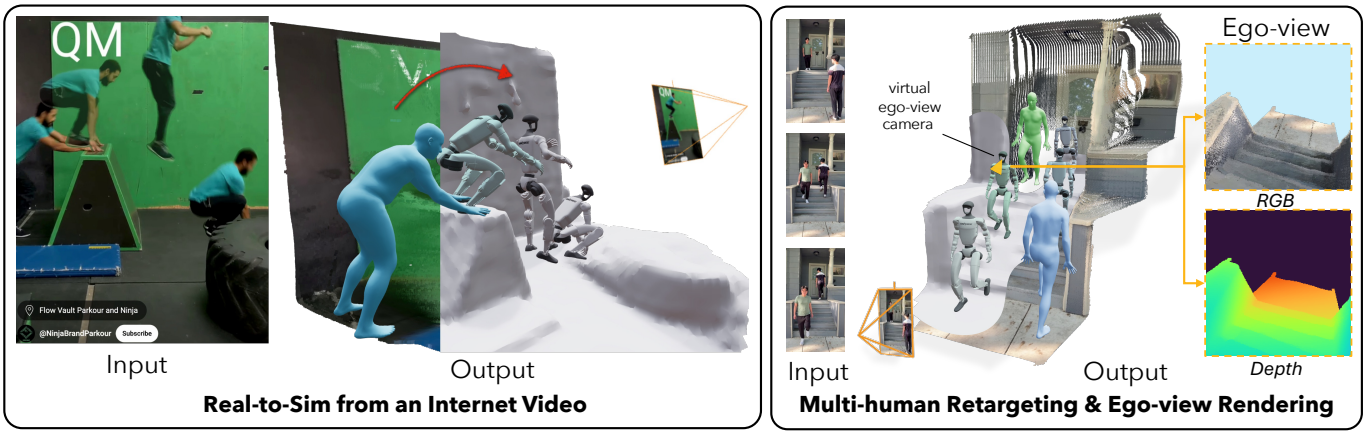

Fig. 4: **Versatile capabilities of our Real-to-Sim pipeline.** VideoMimic enables (i) robust tracking of Internet videos with challenging motion and diverse environments, (ii) simultaneous reconstruction and retargeting of multiple humans, and (iii) ego-view RGB-D rendering for embodied perception—though not used in our current policy, it highlights the framework's broader applicability across inputs and tasks.

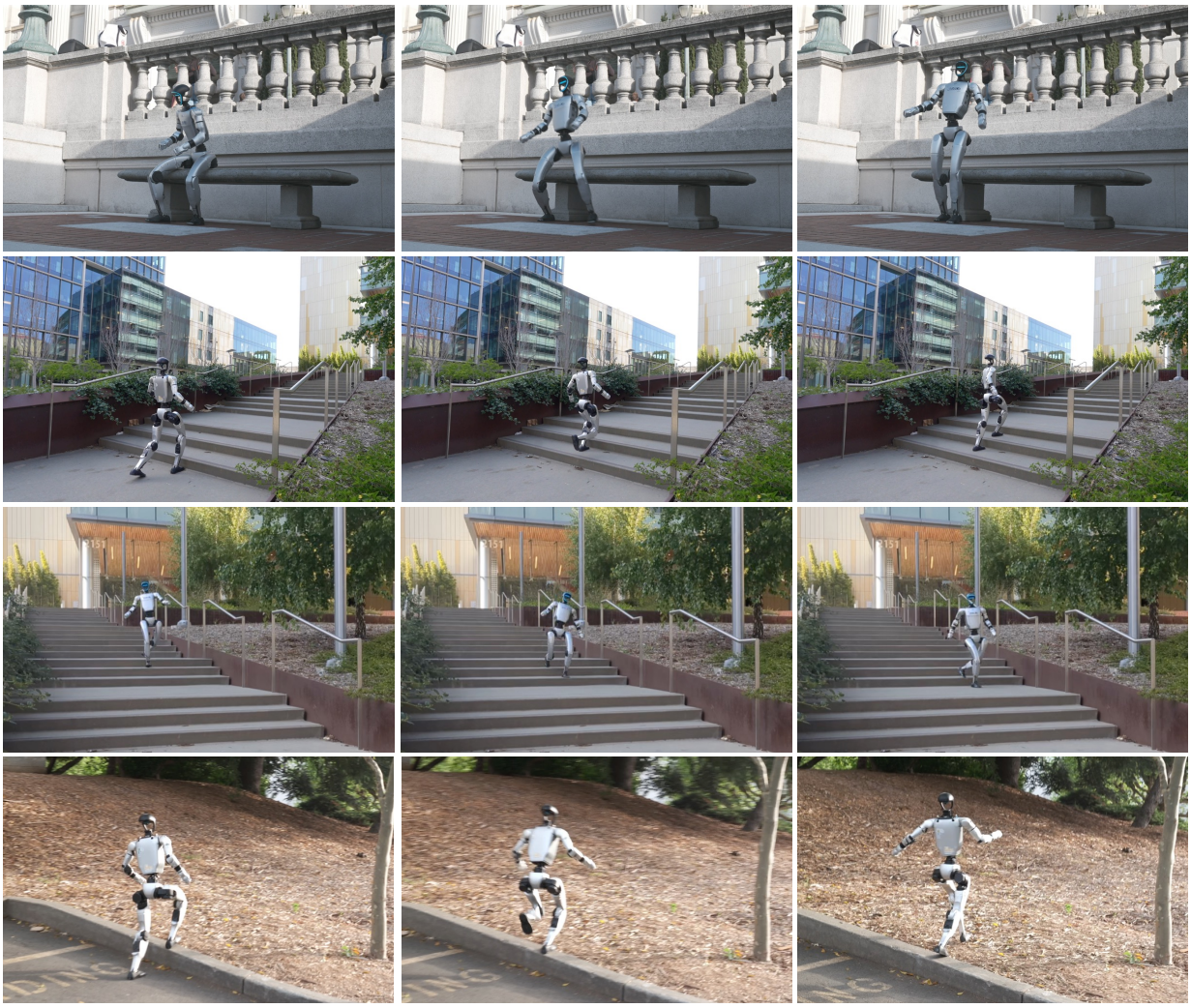

Fig. 5: **The policy performing various skills on the real robot:** traversing complex terrain, standing, and sitting. All these skills are in a single policy, which decides what to do based on the context of its heightmap and joystick direction input. *Top row*: the policy stands from a seated position after sitting down. *Second row*: the policy walks up a flight of stairs. *Third row*: the policy walks down a flight of stairs. *Bottom row*: the policy walks over a kerb and onto a rough terrain. Please find the video results on our webpage.

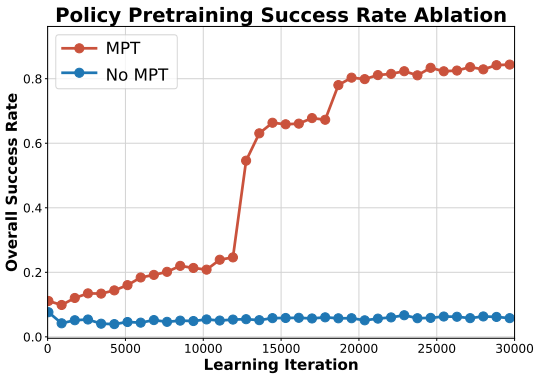

Fig. 6: **Impact of MoCap pre-training (MPT).** Pre-training the policy on motion-capture data facilitates learning on video captures despite noisy references.

initial phases, whereas a policy from scratch may not even be able to learn how to balance. As shown in Figure 6, removing MPT significantly hinders the policy's ability to learn effective behaviors.

B. Real-world Deployment

Setup. We deploy our controller on a 23-DoF *Unitree G1* humanoid and run it onboard at 50 Hz. Following [60], we set relatively low joint gains, $K_p = 75$, to avoid excessively fast or overly stiff behaviour—which helps to avoid excessively violent contact when the robot makes heavy contact with objects such as chairs or stairs. Height-maps are computed in real time using Fast-lio2 [52] and probabilistic terrain mapping [3, 22]. We feed joystick targets from a human operator. Including policy running, all operations are run onboard. We found two critical ingredients for successful motion deployment through iterative sim-to-real trials: (i) relaxing the episode-termination tolerances with respect to the reference motion, and (ii) injecting realistic physics perturbations during training.

Real-world evaluation. Figure 5 and the accompanying video showcase the policy executing a wide range of whole-body behaviors on the Unitree G1. Without any task-specific tuning, the same network—driven only by proprioception and a noisy LiDAR height-map that provides a full 360° view around the torso—climbs and descends indoor and outdoor staircases, traverses steep earthen slopes and rough vegetation, and reliably sits down on or stands up from chairs and benches. The controller is surprisingly resilient: after unexpected foot slides while descending stairs, it recovers by momentarily hopping on a single leg before regaining nominal gait.

To the best of our knowledge, this is the *first* real-world deployment of a *context-aware* humanoid policy learned from monocular human videos, jointly demonstrating perceptive locomotion and environment-prompted whole-body skills such as sitting, standing, and climbing stairs. Additional qualitative results are available on the project webpage.

VI. Limitations

Our pipeline delivers encouraging real-world results, yet several practical weaknesses remain.

Reconstruction. Monocular 4D human–scene recovery is still brittle in the wild. Camera pose drift in MegaSaM often yields duplicate "ghost" layers of the same surface. Due to its inability to refine the dynamic points, the dynamic points from the person are mistakenly fused into the static point cloud or inaccurately placed (e.g., feet buried beneath the environment). In particular, we found that MegaSaM performs poorly on images with low texture. Depth filtering and spatio-temporal subsampling remove many outlier points, but aggressive thresholds leave holes that hinder meshing. NKSR mitigates noise, yet may oversmooth fine geometry (e.g., narrow stair treads); such high-frequency details are crucial for robot control, and we discard videos where these details are missing after reconstruction. Also, during point-to-mesh conversion, spiky artifacts may appear due to stray points.

Retargeting. The kinematic optimizer assumes every reference pose can be made feasible once scaled to the robot. In cluttered scenes, this is not always true, and conflicting costs—strict foot-contact matching versus collision avoidance—can trap the solver in poor local minima that the RL controller must subsequently "clean up."

Sensing and policy input. At test time, the controller receives only proprioception and an 11×11 LiDAR height-map. This coarse grid is adequate for terrain and chairs but lacks the resolution for precise contacts, manipulation, or reasoning about overhanging obstacles. Incorporating richer perceptual inputs—such as RGB-D data or learned occupancy grids—would likely broaden the method's applicability and improve its semantic understanding of the environment.

Simulation fidelity. We assume the scene can be represented as a single rigid mesh. Scaling to articulated or deformable objects will require more expressive simulators and object-level reconstruction pipelines—open problems for future work.

VII. Conclusion

We introduced VideoMimic, a real-to-sim-to-real pipeline that converts everyday human videos into environment-conditioned control policies for humanoids. The system (i) reconstructs humans and surrounding geometry from monocular clips, (ii) retargets the motion to a kinematically feasible humanoid, and (iii) uses the recovered scene as task terrain for dynamics-aware RL. The result is a *single* policy that delivers robust, repeatable contextual control—e.g., stair ascents/descents and chair sit-stand—all driven only by the environment geometry and a root direction command. VideoMimic offers a scalable path for teaching humanoids contextual skills directly from videos. We expect future work to extend the system to richer human–environment interactions, multi-modal sensor-based context learning, and multi-agent behavior modeling, among other directions.

References

[1] Ananye Agarwal, Ashish Kumar, Jitendra Malik, and Deepak Pathak. Legged locomotion in challenging terrains using egocentric vision, 2022. URL https://arxiv.org/abs/2211.07638.

[2] Yudi Dai, YiTai Lin, XiPing Lin, Chenglu Wen, Lan Xu, Hongwei Yi, Siqi Shen, Yuexin Ma, and Cheng Wang. Sloper4d: A scene-aware dataset for global 4d human pose estimation in urban environments. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 682–692, 2023.

[3] Péter Fankhauser, Michael Bloesch, and Marco Hutter. Probabilistic terrain mapping for mobile robots with uncertain localization. *IEEE Robotics and Automation Letters (RA-L)*, 3(4):3019–3026, 2018. doi: 10.1109/LRA.2018.2849506.

[4] Félix G. Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)*, 39(4), 2020.

[5] Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. Synthesizing physical character-scene interactions. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–9, 2023.

[6] Tairan He, Zhengyi Luo, Wenli Xiao, Chong Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Learning human-to-humanoid real-time whole-body teleoperation, 2024. URL https://arxiv.org/abs/2403.04436.

[7] Tairan He, Jiawei Gao, Wenli Xiao, Yuanhang Zhang, Zi Wang, Jiashun Wang, Zhengyi Luo, Guanqi He, Nikhil Sobanbab, Chaoyi Pan, Zeji Yi, Guannan Qu, Kris Kitani, Jessica Hodgins, Linxi "Jim" Fan, Yuke Zhu, Changliu Liu, and Guanya Shi. Asap: Aligning simulation and real-world physics for learning agile humanoid whole-body skills, 2025. URL https://arxiv.org/abs/2502.01143.

[8] David Hoeller, Nikita Rudin, Dhionis Sako, and Marco Hutter. Anymal parkour: Learning agile navigation for quadrupedal robots, 2023. URL https://arxiv.org/abs/2306.14874.

[9] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 13274–13285, June 2022.

[10] Jiahui Huang, Zan Gojcic, Matan Atzmon, Or Litany, Sanja Fidler, and Francis Williams. Neural kernel surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4369–4379, 2023.

[11] Jemin Hwangbo, Joonho Lee, and Marco Hutter. Per-contact iteration method for solving contact dynamics. *IEEE Robotics and Automation Letters*, 3(2):895–902, 2018. URL www.raisim.com.

[12] Jemin Hwangbo, Joonho Lee, Alexey Dosovitskiy, Dario Bellicoso, Vassilios Tsounis, Vladlen Koltun, and Marco Hutter. Learning agile and dynamic motor skills for legged robots. *CoRR*, abs/1901.08652, 2019. URL http://arxiv.org/abs/1901.08652.

[13] Mazeyu Ji, Xuanbin Peng, Fangchen Liu, Jialong Li, Ge Yang, Xuxin Cheng, and Xiaolong Wang. Exbody2: Advanced expressive humanoid whole-body control, 2025. URL https://arxiv.org/abs/2412.13196.

[14] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018.

[15] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020.

[16] Muhammed Kocabas, Ye Yuan, Pavlo Molchanov, Yunrong Guo, Michael J Black, Otmar Hilliges, Jan Kautz, and Umar Iqbal. Pace: Human and camera motion estimation from in-the-wild videos. In *2024 International Conference on 3D Vision (3DV)*, pages 397–408. IEEE, 2024.

[17] Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. RMA: rapid motor adaptation for legged robots. *CoRR*, abs/2107.04034, 2021. URL https://arxiv.org/abs/2107.04034.

[18] Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning quadrupedal locomotion over challenging terrain. *Sci. Robotics*, 5(47):5986, 2020. doi: 10.1126/scirobotics.abc5986. URL https://doi.org/10.1126/scirobotics.abc5986.

[19] Jiefeng Li, Siyuan Bian, Chao Xu, Gang Liu, Gang Yu, and Cewu Lu. D &d: Learning human dynamics from dynamic camera. In *European Conference on Computer Vision*, pages 479–496. Springer, 2022.

[20] Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. Megasam: Accurate, fast, and robust structure and motion from casual dynamic videos. *arXiv preprint arXiv:2412.04463*, 2024.

[21] Zhizheng Liu, Joe Lin, Wayne Wu, and Bolei Zhou. Joint optimization for 4d human-scene reconstruction in the wild. *arXiv preprint arXiv:2501.02158*, 2025.

[22] Junfeng Long, Junli Ren, Moji Shi, Zirui Wang, Tao Huang, Ping Luo, and Jiangmiao Pang. Learning humanoid locomotion with perceptive internal model. *arXiv preprint arXiv:2411.14386*, 2024.

[23] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–

866. 2023.

[24] Tyler Ga Wei Lum, Martin Matak, Viktor Makoviychuk, Ankur Handa, Arthur Allshire, Tucker Hermans, Nathan D. Ratliff, and Karl Van Wyk. Dextrah-g: Pixels-to-action dexterous arm-hand grasping with geometric fabrics, 2024. URL https://arxiv.org/abs/2407.02274.

[25] Zhengyi Luo, Jinkun Cao, Josh Merel, Alexander Winkler, Jing Huang, Kris M. Kitani, and Weipeng Xu. Universal humanoid motion representations for physics-based control. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=OrOd8PxOO2.

[26] Diogo C Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5137–5146, 2018.

[27] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, and Gavriel State. Isaac gym: High performance gpu-based physics simulation for robot learning. *CoRR*, abs/2108.10470, 2021. URL https://arxiv.org/abs/2108.10470.

[28] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3d gaussian avatar. In *European Conference on Computer Vision*, pages 19–35. Springer, 2024.

[29] Lea Müller, Hongsuk Choi, Anthony Zhang, Brent Yi, Jitendra Malik, and Angjoo Kanazawa. Reconstructing people, places, and cameras. *arXiv preprint arXiv:2412.17806*, 2024.

[30] Liang Pan, Zeshi Yang, Zhiyang Dou, Wenjia Wang, Buzhen Huang, Bo Dai, Taku Komura, and Jingbo Wang. Tokenhsi: Unified synthesis of physical human-scene interactions through task tokenization, 2025. URL https://arxiv.org/abs/2503.19901.

[31] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9054–9063, 2021.

[32] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. DeepMimic. *ACM Transactions on Graphics*, 37(4):1–14, jul 2018. doi: 10.1145/3197517.3201311. URL https://doi.org/10.1145%2F3197517.3201311.

[33] Xue Bin Peng, Angjoo Kanazawa, Jitendra Malik, Pieter Abbeel, and Sergey Levine. SFV: reinforcement learning of physical skills from videos. *CoRR*, abs/1810.03599, 2018. URL http://arxiv.org/abs/1810.03599.

[34] Ilija Radosavovic, Tete Xiao, Bike Zhang, Trevor Darrell, Jitendra Malik, and Koushil Sreenath. Real-world humanoid locomotion with reinforcement learning, 2023. URL https://arxiv.org/abs/2303.03381.

[35] Ilija Radosavovic, Sarthak Kamat, Trevor Darrell, and Jitendra Malik. Learning humanoid locomotion over challenging terrain. *arXiv preprint arXiv:2410.03654*, 2024.

[36] Ilija Radosavovic, Bike Zhang, Baifeng Shi, Jathushan Rajasegaran, Sarthak Kamat, Trevor Darrell, Koushil Sreenath, and Jitendra Malik. Humanoid locomotion as next token prediction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[37] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people by predicting 3d appearance, location and pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2740–2749, 2022.

[38] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, Christoph Feichtenhofer, and Jitendra Malik. On the benefits of 3d pose and tracking for human action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 640–649, 2023.

[39] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL https://arxiv.org/abs/2408.00714.

[40] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024.

[41] Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. No-regret reductions for imitation learning and structured prediction. *CoRR*, abs/1011.0686, 2010. URL http://arxiv.org/abs/1011.0686.

[42] Nikita Rudin, David Hoeller, Philipp Reist, and Marco Hutter. Learning to walk in minutes using massively parallel deep reinforcement learning. In Aleksandra Faust, David Hsu, and Gerhard Neumann, editors, *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 91–100. PMLR, 08–11 Nov 2022. URL https://proceedings.mlr.press/v164/rudin22a.html.

[43] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL http://arxiv.org/abs/1707.06347.

[44] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: Reconstructing world-grounded humans with accurate 3D motion. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[45] Himanshu Gaurav Singh, Antonio Loquercio, Carmelo Sferrazza, Jane Wu, Haozhi Qi, Pieter Abbeel, and Jitendra Malik. Hand-object interaction pretraining from videos, 2024. URL https://arxiv.org/abs/2409.08273.

[46] Chen Tessler, Yunrong Guo, Ofir Nabati, Gal Chechik, and Xue Bin Peng. Maskedmimic: Unified physics-based character control through masked motion inpainting. *ACM Transactions on Graphics (TOG)*, 2024.

[47] Nicolas Ugrinovic, Boxiao Pan, Georgios Pavlakos, Despoina Paschalidou, Bokui Shen, Jordi Sanchez-Riera, Francesc Moreno-Noguer, and Leonidas Guibas. Multiphys: multi-person physics-aware 3d motion estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2331–2340, 2024.

[48] Alexander Veicht, Paul-Edouard Sarlin, Philipp Lindenberger, and Marc Pollefeys. Geocalib: Learning single-image calibration with geometric optimization. In *European Conference on Computer Vision*, pages 1–20. Springer, 2024.

[49] Huayi Wang, Zirui Wang, Junli Ren, Qingwei Ben, Tao Huang, Weinan Zhang, and Jiangmiao Pang. Beamdojo: Learning agile humanoid locomotion on sparse footholds. In *Robotics: Science and Systems (RSS)*, 2025.

[50] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. *arXiv preprint arXiv:2501.12387*, 2025.

[51] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. *arXiv preprint arXiv:2403.17346*, 2024.

[52] Wei Xu, Yixi Cai, Dongjiao He, Jiarong Lin, and Fu Zhang. Fast-lio2: Fast direct lidar-inertial odometry, 2021. URL https://arxiv.org/abs/2107.06829.

[53] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584, 2022.

[54] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. 2023.

[55] Brent Yi, Vickie Ye, Maya Zheng, Yunqi Li, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. Estimating body and hand motion in an ego-sensed world. *arXiv preprint arXiv:2410.03665*, 2024.

[56] Ri Yu, Hwangpil Park, and Jehee Lee. Human dynamics from monocular video with dynamic camera movements. *ACM Transactions on Graphics (TOG)*, 40 (6):1–14, 2021.

[57] Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. Simpoe: Simulated character control for 3d human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7159–7169, 2021.

[58] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11038–11049, 2022.

[59] Ye Yuan, Viktor Makoviychuk, Y Guo, S Fidler, XB Peng, and K Fatahalian. Learning physically simulated tennis skills from broadcast videos. *ACM Trans. Graph*, 42(4), 2023.

[60] Kevin Zakka, Baruch Tabanpour, Qiayuan Liao, Mustafa Haiderbhai, Samuel Holt, Jing Yuan Luo, Arthur Allshire, Erik Frey, Koushil Sreenath, Lueder A. Kahrs, Carmelo Sferrazza, Yuval Tassa, and Pieter Abbeel. Mujoco playground, 2025. URL https://arxiv.org/abs/2502.08844.

[61] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024.

[62] Siwei Zhang, Yan Zhang, Federica Bogo, Marc Pollefeys, and Siyu Tang. Learning motion priors for 4d human body capture in 3d scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11343–11353, 2021.

[63] Ziwen Zhuang, Shenzhe Yao, and Hang Zhao. Humanoid parkour learning, 2024. URL https://arxiv.org/abs/2406.10759.