HSImul3R: Physics-in-the-Loop Reconstruction of Simulation-Ready Human–Scene Interactions

Yukang Cao[1], Haozhe Xie[1], Fangzhou Hong[1], Long Zhuo[3], Zhaoxi Chen[1], Liang Pan[2], and Ziwei Liu[1, ✉]

[1] S-Lab, Nanyang Technological University [2] ACE Robotics [3] Shanghai AI Laboratory

✉ Corresponding author

https://yukangcao.github.io/HSImul3R/

Fig. 1: Examples of real-world humanoid robotic deployment. Given casual captures, our approach achieves simulation-ready 3D reconstruction of human–scene interactions by refining the human motions and scene geometry via a physically-grounded bi-directional optimization pipeline. Our optimized human motions can be seamlessly transferred and deployed in humanoid robotics.

Abstract. We present **HSImul3R**[1], a unified framework for simulation-ready 3D reconstruction of human-scene interactions (HSI) from casual captures, including sparse-view images and monocular videos. Existing methods suffer from a *perception-simulation gap*: visually plausible reconstructions often violate physical constraints, leading to instability in physics engines and failure in embodied AI applications. To bridge this gap, we introduce a **physically-grounded bi-directional optimization pipeline** that treats the physics simulator as an active supervisor to jointly refine human dynamics and scene geometry. In the forward direction, we employ *Scene-targeted Reinforcement Learning* to optimize human motion under dual supervision of motion fidelity and contact stability. In the reverse direction, we propose *Direct Simulation Reward Optimization*, which leverages simulation feedback on gravitational stability and interaction success to refine scene geometry. We further present

[1] Pronunciation: / ˈsɪmjulə(r) /

HSIBench, a new benchmark with diverse objects and interaction scenarios. Extensive experiments demonstrate that HSImul3R produces the first stable, simulation-ready HSI reconstructions and can be directly deployed to real-world humanoid robots.

Keywords: Human-Scene-Interaction · Physical Simulation · Humanoid Embodied AI

1 Introduction

Embodied artificial intelligence aims to integrate intelligent agents into daily life through physically grounded systems. Unlike disembodied models [8, 10, 11, 83] limited to virtual domains, embodied AI [79, 91, 92, 94] learns transferable motions that enable perception, reasoning, and action in real-world environments. A key challenge is modeling humanoid–scene interactions, requiring understanding of human motion, spatial layouts, and interaction stability. Reconstructing human–scene interactions (HSI) [4, 40, 82] from images or videos provides high-fidelity supervision and enables scalable, simulation-ready datasets, helping bridge passive observation and active robotic deployment.

Current methods suffer from a *perception–simulation gap*, where visually plausible reconstructions violate physical constraints and fail in embodied AI applications. This gap largely stems from the fragmented modeling of humans and environments, as existing approaches rarely capture their explicit physical coupling and instead fall into three separate directions: **1) 3D scene reconstruction** (*e.g*, NeRF [51], Gaussian Splatting [28], DUSt3R [72]), which prioritizes environment geometry while largely ignoring human dynamics. **2) Human motion estimation** [9, 30, 57, 70], which achieves robustness under occlusion but reconstructs motion in isolation, without modeling physical contact or environmental constraints. **3) Interaction modeling** [13, 15, 25, 55, 89], typically based on SMPL-driven HSI datasets [5, 24, 41] that remain limited in scale, diversity, and physical validation. Recent unified frameworks (*e.g*, HOSNeRF [36], HSfM [52]) attempt joint modeling but optimize mainly in the 2D image space, prioritizing visual alignment over geometric and physical validity. Consequently, the resulting reconstructions lack metric and contact fidelity, making them unsuitable for simulation and preserving the gap between visual realism and embodied deployment.

To close this gap, we introduce HSImul3R, a simulation-ready Human–Scene Interaction 3D reconstruction framework that formulates reconstruction as a bidirectional physics-aware optimization problem. A physics simulator acts as an active supervisor, enabling closed-loop refinement between human motion and scene geometry. HSImul3R operates along two complementary directions. **Forward optimization** refines human motion under fixed scene geometry. After establishing metric-consistent human–scene alignment with structural priors from image-to-3D generative models, we integrate the reconstruction into the simulator and perform scene-targeted reinforcement learning. Motion is optimized using physically grounded signals, including keypoint tracking consistency and

Fig. 2: Examples of HSIBench and corresponding simulation results by HSImul3R. Our approach enables simulation-ready 3D reconstruction of human–scene interactions from casual captures. In addition, we collect HSIBench, a dataset comprising 16-view synchronized captures of diverse human–scene interactions, covering a wide range of scene objects, human subjects, and motions.

geometric contact constraints, improving interaction stability. **Reverse optimization** refines scene geometry under physically validated motion. To address instability caused by structurally deficient geometry, we introduce Direct Simulation Reward Optimization (DSRO), which leverages simulator-derived rewards to enhance gravitational stability and interaction feasibility.

To support the training and benchmarking of this framework, we collect **HSIBench**, a new dataset comprising 19 objects and over 50 motion sequences recorded by two male and one female participants, totaling 300 unique interaction instances. An overview of HSIBench and simulation results of our method is provided in Fig. 2.

We conduct extensive experiments to evaluate HSImul3R against state-of-the-art baselines in terms of simulation stability, post-simulation human motion quality, and improvements in image-to-3D generation through DSRO fine-tuning. Experimental results demonstrate that HSImul3R is the first approach to achieve stable, simulation-ready reconstructions of human–scene interactions, offering robust performance across diverse scenarios and significantly outperforming existing techniques. Finally, we demonstrate the real-world utility of our framework by (1) retargeting the refined motions to a Unitree humanoid robot, and (2) training a whole-body motion tracking policy for physical deployment. Examples of real-world deployment are presented in Fig. 1.

2 Related Works

3D Scene Reconstruction. Early approaches are dominated by geometry-based methods, such as structure-from-motion [65] and multi-view stereo [66], which estimate camera poses and dense geometry from multiple views. With the rise of deep learning, data-driven approaches emerge, including monocular depth prediction [87, 88] and learning-based multi-view stereo [22], enabling reconstruction from sparse or unstructured imagery. Other works adopt explicit 3D representations such as voxels [37,68], point clouds [14,80], and meshes [54], often optimized through differentiable rendering. More recently, implicit neural representations, such as signed distance functions [56], occupancy fields [6], neural radiance fields [33,76], and explicit but differentiable formulations like 3D Gaussian Splatting [28, 77], become central to high-quality scene modeling. Beyond static reconstruction, dynamic scene modeling [78, 86] expands these methods to time-varying environments. In parallel, recent works such as Dust3R [72] and VGGT [71] introduce pre-trained transformers that enable end-to-end 3D reconstruction directly from uncalibrated and unlocalized images, eliminating the need for expensive post-optimization.

Physically-sounded Modeling. Recent works have sought to embed physical soundness into modeling, which can be broadly categorized into three paradigms. Physics-constrained and physics-integrated generation methods unify simulation and content creation by leveraging simulation-derived losses or physical priors. For example, PhyRecon [53] ensures stable scene reconstruction, Atlas3D [12] and BrickGPT [61] produce self-supporting structures, and DSO [31] or PhysDeepSDF [50] align generators with simulation feedback. PhysGaussian [81] evolves Gaussian splats via continuum mechanics, while PhyCAGE [85], VR-GS [26], and GASP [7] optimize assets through MPM; PAC/iPAC-NeRF [27,32] jointly learn geometry and physical parameters to bridge reconstruction and simulation. This approach also extends to interactive contexts: PhyScene [90] generates simulation-ready environments, PhysPart [42] models functional parts for robotics and fabrication, and DreMa [3] produces manipulable, physics-grounded world models.

Human Simulation Imitating. Recent advances in physics-based humanoid simulation fall into three directions. Robust motion imitation builds on RL frameworks such as DeepMimic [58] and AMP [60], extended by PHC [45] for long-horizon resilience and DiffMimic [63] with differentiable physics. More recent methods leverage human demonstrations for adaptive whole-body imitation, including locomotion and manipulation, as in HumanPlus [16], TWIST [93], and SFV [59]. Generalizable control is advanced by PULSE [44], which provides compact latent spaces for versatile skills, HOVER [19], which unifies multiple control modes, and diffusion-based frameworks such as CLoSD [69] and InsActor [64], which integrate generative planning with physics-based execution for multi-task behaviors. Interactive skills cover dynamic human-object interactions and complex benchmarks: PhysHOI [73] and OmniGrasp [43] enable dexterous manip-

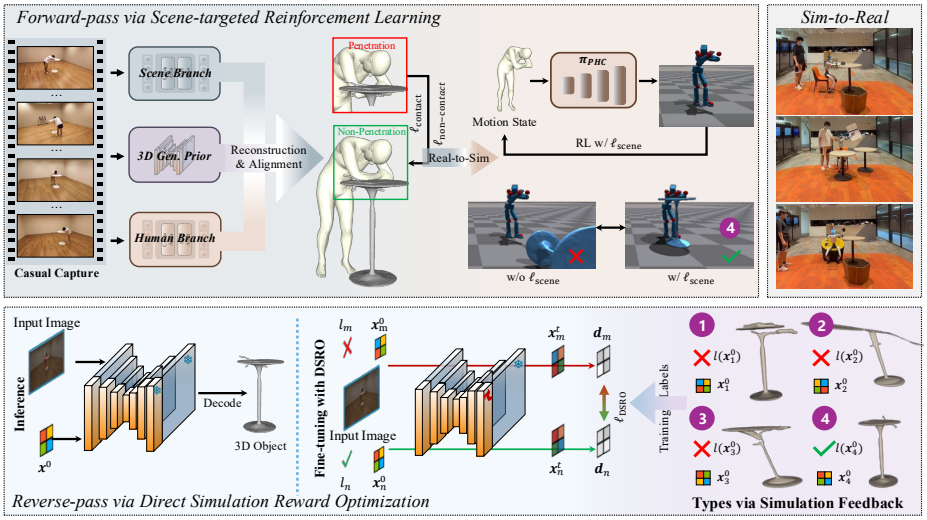

Fig. 3: Overview of HSImul3R. Given casual captures as inputs, we achieve simulation-ready reconstruction of human–scene interactions via a physics-in-the-loop optimization pipeline. We first propose to inject an 3D explicit generative prior into the reconstruction pipeline to achieve better alignment between human and scene. Then, **(1)** in the forward-pass, we propose a scene-targeted reinforcement learning that optimize the human motion to achieve interaction stability within the simulator, **(2)** in the reverse-pass, we introduce a direct simulation reward optimization (DSRO) to refine the scene geometry via simulation feedback regarding the stability. Specifically, we define the 4 types regarding the feedback. Type 1: objects not stabilizing under gravity; Type 2: objects failing to stabilize during human interaction; Type 3: objects stabilizing but without meaningful interaction; Type 4: objects with stable interaction.

ulation, SMPLOlympics [46] and HumanoidOlympics [47] provide sports environments, Half-Physics [67] bridges kinematic avatars with physics, ImDy [38] exploits imitation-driven simulation, ASAP [18] improves fidelity by aligning dynamics with demonstration trajectories, BeyondMimic [35] learns a motion tracking policy that could be deployed into humanoid robot, and VideoMimic [1] enables learning such policies from as little as a single monocular video.

3 Our Approach

As illustrated in Fig. 3, HSImul3R can reconstruct simulation-ready human-scene-interactions (HSI) from casual captures. To achieve this, we first reconstruct both human motion and scene geometry, subsequently aligning them through an explicit 3D structural prior derived from image-to-3D generative models [23] (Sec. 3.2). Following this reconstruction, we introduce a physically-grounded bi-directional optimization pipeline. This process consists of a forward-pass optimization, which employs a proposed scene-targeted reinforcement learn-

ing scheme to refine human motions (Sec. 3.3), followed by a reverse-pass optimization that leverages simulator feedback regarding physical stability to rectify the structural correctness of the scene geometry (Sec. 3.4). For simplicity of the illustration, we first focus on the setting of $J = 4$ uncalibrated sparse-view inputs in the following sections and then discuss its extension to monocular videos in Sec. 3.5. Preliminaries underlying our methodology are provided in Sec. 3.1.

3.1 Preliminaries

DUSt3R. Recently, DUSt3R [72] introduced a framework for 3D reconstruction that regresses point maps and employs a global alignment strategy to jointly predict depth maps and camera poses. Specifically, given a set of input images $\mathbf{I} = I_0, I_1, ..., I_J$, DUSt3R applies a ViT-based network that takes a pair of image frames I_n, I_m $(n, m \in [0, J])$ to estimate the corresponding point maps $P_n^e, P_m^e \in \mathbb{R}^{H\times W\times 3}$ with respect to the coordinate system of frame n, along with confidence maps $C_n^e, C_m^e \in \mathbb{R}^{H\times W}$. Here, $e = (m, n)$ denotes the selected image pair. Aggregating point maps and confidence maps across selected pairs, DUSt3R builds a connectivity graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ corresponds to the N images and $\mathcal{E}$ to the chosen image pairs e.

After collecting all pairwise point maps, DUSt3R performs a global alignment optimization to recover the depth maps $\mathbf{D} = D_0, D_1, ..., D_J$ and camera poses $\pi_0, \pi_1, ..., \pi_J$:

$$\arg\min_{D,\pi,\sigma} \sum_{e\in\mathcal{E}} \sum_{n\in e} C_n^e ||D_n - \sigma_e \cdot F_e(\pi_n, P_n^e)||_2^2, \tag{1}$$

where $\sigma = \sigma_e, e \in \mathcal{E}$ denotes the edge-wise scale factors, and $F_e(\pi_n, P_n^e)$ projects the predicted point map P_n^e to view n under camera pose π_n to produce the corresponding depth. This objective enforces geometric alignment across the input frame pairs, ensuring cross-view consistency in the estimated depth maps after the optimization. However, DUSt3R struggles with human subjects and frequently produces reconstruction artifacts such as unrealistic/incorrect structures and non-watertight topologies. Such defects make the reconstructed environments unreliable for stable simulation and downstream embodied AI tasks.

3.2 Human-Scene Interaction Reconstruction and Alignment

The first stage of our pipeline involves the independent reconstruction of the static scene geometry and the dynamic human motion from uncalibrated captures. We adopt DUSt3R to recover the 3D structure of the environment. For human motion estimation, we first utilize SAM2 [62] to detect and associate individuals across frames, generating precise masks and identity tracks. Following this, we employ 4DHumans [17] and ViTPose [84] to extract the initial 3D SMPL-based motion sequences and 2D keypoints (J_{2D}), respectively. As the initial human and scene reconstructions may reside in disparate coordinate spaces, we perform a joint optimization to unify them [52]. This is achieved through: (1) human-centric bundle adjustment guided by the 2D keypoints J_{2D}, and (2)

global human-scene alignment, which minimizes the reprojection error between the ViTPose-detected keypoints and the projected 3D SMPL joints to ensure spatial consistency.

Alignment via Explicit 3D Structural Prior. Despite the initial alignment listed above, two critical issues often persist: (1) the reconstructed scene geometry frequently contains structural artifacts, such as disconnected components, missing surfaces, or non-watertight topologies; and (2) the human-scene alignment relies solely on 2D projection-based supervision, which lacks 3D geometric awareness and is vulnerable to occlusions. These deficiencies inevitably lead to physical instability and drifting within the physics simulator. To resolve these challenges, we leverage 3D structural priors from pre-trained generative models to rectify the scene's geometry and enforce more robust interaction constraints.

Concretely, for each object within the scene, we automatically identify the input image $I_n, n \in [1, J]$ in which the object is most prominently featured and employ SAM [29] to extract its segmentation mask. This view is then processed by a pre-trained image-to-3D generative model [23] to synthesize a high-fidelity 3D representation with better structural accuracy:

$$\mathcal{R}_{\text{scene}} := \{\texttt{MIDI}(I_n[M_i]), i \in [0, O]\}, \tag{2}$$

where $\mathcal{R}_{\text{scene}}$ denotes the refined 3D scene and O is the total number of objects. Note that our framework is flexible and allows for the usage of alternative or future more advanced models as they become available.

Thanks to the injection of 3D structural priors, we are now able to refine the human-scene alignment with 3D explicit constraints. This process is essential because penetration artifacts are particularly problematic in simulation: even minor inconsistencies in 3D space can manifest as severe collisions between body parts and objects, ultimately leading to unstable or failed simulations. To this end, we propose to optimize the position of the recovered human and generated objects []. Specifically, if the object and human are not in contact, we optimize their positions via:

$$\begin{aligned}\ell_{\text{non-contact}} = &\frac{1}{|H_p|} \cdot \sum_{1 \le j \le N_o} ||\mu_i^h - \mu_j^o||_2 \\ &+ \frac{1}{N_o} \cdot \sum_{j=1}^{N_o} \min_{i \in H_p} ||\mu_j^o - \mu_i^h||_2,\end{aligned} \tag{3}$$

where H_P denotes the human body part closest to the object, and N_o is the number of vertices on the object, and μ_j^o and μ_i^h represent the 3D positions of object and human vertices, respectively. When the object is in contact with the human, we instead apply:

$$\ell_{\text{contact}} = \frac{1}{|H_p|} \cdot \sum_{i \in H_p} \max(0, -\delta(\mu_i^h)), \tag{4}$$

where $\delta(\cdot)$ denotes the signed distance function, measuring the penetration depth of the human vertex μ_i^h relative to the object surface.

3.3 Forward-Pass: Scene-Targeted Motion Optimizatiton

Following the initial 3D reconstruction of the human-scene interaction, the next step is to ensure stable dynamics within a physics simulator [48]. A direct approach is to employ motion tracking techniques [45] to retarget the reconstructed human poses onto a humanoid robot. However, directly simulating the raw reconstructions often fails to yield stable interactions (see Fig. 5). In many cases, the humanoid inadvertently displaces nearby objects, leaving them separated from the body and resting independently on the ground. This instability occurs because conventional 3D reconstructions do not account for gravity and interaction forces to verify if poses and object placements are physically realizable.

To address this issue, we introduce a scene-targeted supervision signal to reinforcement-learning-based motion tracking [45]. Specifically, we propose an objective that enforces spatial proximity between the humanoid and relevant scene objects, encouraging physically plausible contact during simulation. This loss is defined as the average Euclidean distance between human contact keypoints k_j^h and their corresponding nearest object surface points μ_i^o.

$$\ell_{\text{scene}} = \frac{1}{N_{\text{contact}} \cdot N_{\text{surf}}} \cdot \sum_{j=1}^{N_{\text{contact}}} \sum_{i=1}^{N_{\text{surf}}} \|\mu_i^o - k_j^h\|_2^2, \tag{5}$$

where N_{contact} is the number of contacts between the human and scene objects, and N_{surf} denotes the number of sampled object surface points within the local contact region.

3.4 Reverse-Pass: Simulator-Guided Object Refinement

Nonetheless, even our forward-pass with scene-targeted reinforcement learning could enhance the simulation stability, we may still observe unsatisfactory stability ratios (see Tab. 1). As presented in Fig. 4, we observe that this problem largely stems from the inconsistent quality of our explicit 3D generative prior, for two main reasons: (1) generated objects often contain structural defects, especially in slender geometries. For example, tables or chairs may be missing legs, making them unstable in the simulator even without interaction; and (2) severe occlusion by the human in the input images, which frequently happens, often results in generated objects exhibiting artifacts, such as surface distortions or unwanted bumps. Together, these limitations make it difficult for the humanoid to establish stable and physically plausible contact during simulation.

Direct Simulation Reward Optimization. Inspired by DSO [31], we address this issue by introducing Direct Simulation Reward Optimization (DSRO), a novel approach that leverages physics-based simulation feedback as a supervision signal for refining 3D explicit object generation. Unlike methods that rely on human annotations or 3D ground truth, DSRO directly exploits the outcome of the simulation to assess the physical plausibility of generated objects and their interactions with humans.

Formally, we define the DSRO objective as:

$$\ell_{\text{DSRO}} = -T \cdot \mathbb{E}_{I\sim\mathcal{I},x_0\sim\mathcal{X}_I,t\sim\mu(0,T),x_t\sim q(x_t|x_0)} \\ [w(t)\cdot(1-2\cdot l(x_0))||\epsilon-\epsilon_\theta(x_t,t)||_2^2]], \tag{6}$$

where I denotes an image sampled from the training dataset $\mathcal{I}$, $\mathcal{X}$ corresponds to its generated 3D explicit object, and $l(\cdot)$ encodes the stability feedback obtained from simulation. Crucially, we define the stability $l(x_0)$ based on both gravitional stability and interaction stability:

$$l(x_0) = \begin{cases} 1, & \text{if stable} \\ 0, & \text{otherwise,} \end{cases} \tag{7}$$

where stability is determined according to three criteria: (1) the object must remain upright and physically stable under gravity within the simulator, (2) it must achieve a stable final state for the reconstructed scene, and (3) the interaction must involve actual contact rather than the object resting independently on the ground.

HSIBench. To support the training and benchmarking of this framework, we construct a dedicated benchmark dataset, HSIBench, tailored for human–scene interaction (HSI). The dataset is built by systematically capturing interaction scenarios involving three volunteers (two male and one female) engaging with a diverse set of objects, including eight chairs, three tables, and three sofas. In total, we record 300 distinct HSI cases, with each case captured from 16 different viewpoints to provide rich multi-view supervision. Representative examples are illustrated in the Appendix. We employ multi-view 2DGS reconstruction [21] and SMPL estimation to respectively derive pseudo ground truth for object geometry and human motion, for our quantitative evaluations. For every captured case, we run our full reconstruction and simulation pipeline, as described in Sec. 3.2 and Sec. 3.3, 15 times under different random seeds for each view. This procedure ensures variability in the simulation outcomes and allows us to systematically collect the training signals needed for fine-tuning.

3.5 Extension to Monocular Videos

Thanks to the dynamic nature of the motion tracking techniques [45] employed in our forward-pass (Sec. 3.3), our method can be easily extended to take monocular video as input to produce simulation-ready 4D reconstructions. In this pipeline, we employ MegaSAM [34] and TRAM [74] for scene reconstruction and human motion estimation [1], respectively. We currently assume a static scene in which the human subject is the only moving entity. To achieve dynamic 3D alignment between the human and the scene geometry, we employ SAM2 [62] to extract 2D bounding boxes for both the person and specific scene geometry (*e.g*, tables, chairs). This spatial knowledge allows us to preliminarily identify interactions and achieve accurate 3D alignment throughout the sequence.

Table 1: Quantitative comparison regarding human-scene-interaction reconstruction and simulation. We assess (1) post-simulation interaction stability, (2) human–scene penetration in the 3D reconstruction, and (3) changes in the human motions after simulation. Our method consistently outperforms the baseline method (HSfM) and different variants.

Method	Stability-HSI (%) ↑			Scene Penetration	Human Motion Quality	
	Easy	Medium	Hard	SP-3D (%) ↓	W-MPJPE ↓	PA-MPJPE ↓
HSfM [52]	10.52	4.50	2.66	69.51	5.02	2.79
V1	13.96	8.81	4.17	77.12	6.18	3.20
V2	39.56	22.71	7.05	-	4.91	2.71
V3	42.57	23.84	10.18	-	4.60	2.42
V4	29.56	16.62	5.17	-	4.57	2.39
Ours	**53.68**	**30.56**	**13.92**	**22.9**	**4.09**	**2.17**

Table 2: Quantitative comparison regarding image-to-3D generation quality. Stability-HSI' denotes the ratio of simulations in which the human–scene interaction remains stable, while "Stability-Gravity"(SG) refers to placing the object alone in the simulator and evaluating whether it can stand stably under gravity. Note that we fine-tune DSO on pre-trained MIDI model for fair comparison.

Method	Stability-HSI (%) ↑			SG (%) ↑	CD ↓	F-Score ↑
	Easy	Medium	Hard			
MIDI [23]	29.56	16.62	5.17	79.19	0.198	81.95
DSO* [31]	38.75	25.91	7.88	87.23	0.191	86.26
Ours	**53.68**	**30.56**	**13.92**	**91.50**	**0.173**	**88.25**

4 Experiments

We evaluate HSImul3R across three dimensions: reconstruction fidelity, simulation stability, and the impact of fine-tuning with the proposed DSRO. We also benchmark against existing methods and perform ablation studies to assess the contribution of each component.

Implementation Details. Our approach is developed on top of HSfM [52] and PHC [45]. For training, we adopt AdamW [39] as the optimizer and fine-tune the pre-trained MIDI model using LoRA [20]. Specifically, we set the LoRA rank to 64, use a batch size of 1, and a learning rate of 10^{-5}. The model is trained for a total of 3000 steps on four NVIDIA A100 GPUs.

Baseline Methods. Since our method presents the first approach for simulation-ready reconstruction of human–scene interactions from uncalibrated sparse-view inputs, we primarily compare its performance against HSfM [52], which is the first and only technique to reconstruct 3D HSI under sparse-view settings. Additionally, considering that there is no other dedicated method existing for this

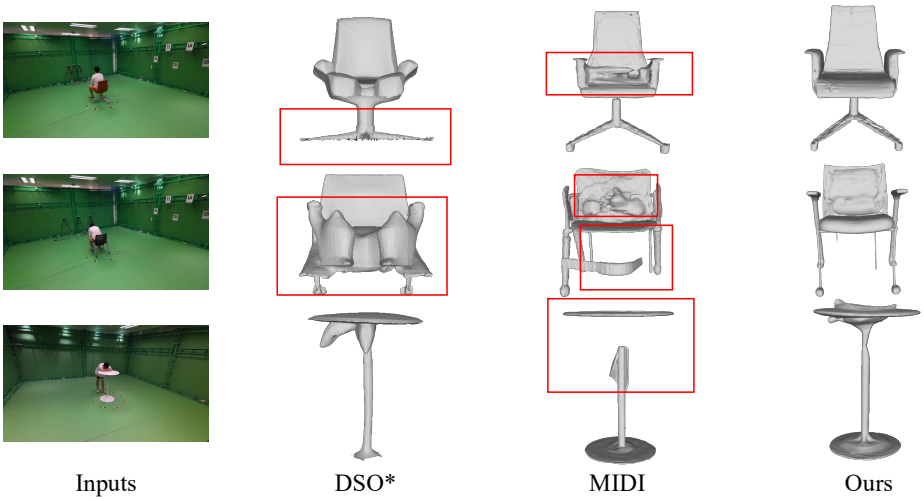

Fig. 4: Qualitative comparison regarding image-to-3D object reconstruction. Our method enhances the object's geometric structure while reducing surface "bumps" that may negatively impact human interaction.

task, we further compare with various alternatives for better evaluations: **(V1 - HSfM w/ watertight 3D scene)** a simple baseline that integrates HSfM with MIDI [23] and feeds the resulting reconstruction into the simulator; **(V2 - Ours w/o bi-directional optimization)** using the reconstruction from Sec. 3.2 directly in the simulator without applying the scene-targeted distance minimization of Eq. 5; **(V3 - Ours w/ alternative Eq. 5)** Replacing our object-surface distance computation with a center-point distance following CLoSD [69]; **(V4 - Ours w/o reverse-pass)** Obtain the simulated reconstruction directly via Sec. 3.2 and Sec. 3.3 without fine-tuning the generative model using simulation feedback via the proposed DSRO.

We also compare with the MIDI [23] and DSO [31] in terms of the geometric quality of the generated scene objects, as well as stability under both gravity-only and HSI scenarios. For fairness, we fine-tune DSO on the pre-trained MIDI model rather than its originally used TRELLIS [75] model.

Evaluation Metrics. We first evaluate the penetration ratio (SP-3D) in the reconstructed 3D HSI scenes. Next, we assess the stability of simulated human–scene interactions using the metric "Stability-HSI", which considers three factors: (1) object stability under gravity, (2) whether the HSI scene reaches a stable state in the simulator, and (3) whether the final state preserves meaningful human–scene interactions. Finally, we evaluate the quality of simulated human motion by extracting it from the final state and comparing it to the ground truth. Following HSfM, we report W-MPJPE for accuracy in the world coordinate system and PA-MPJPE for local pose precision.

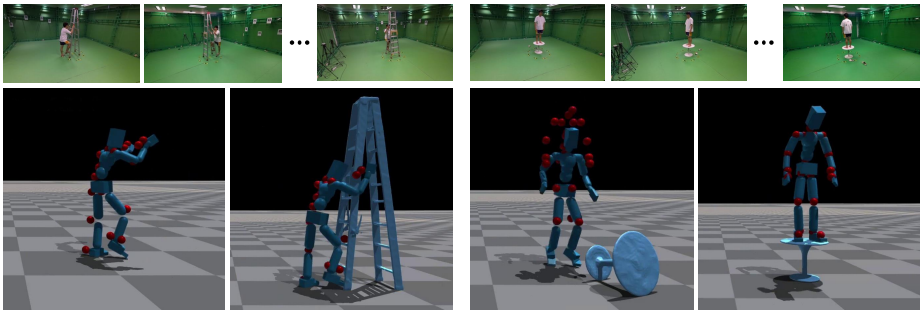

Fig. 5: Qualitative comparisons with HSfM [52]. Due to challenges such as (1) penetration issues and (2) inaccurate scene-object structures with geometric distortions, HSfM often struggles to achieve stable interactions in the simulator, frequently leading to unintended object displacement.

For reconstructed 3D scene objects, we measure geometric quality using Chamfer Distance and F-Score, while physical plausibility is evaluated through "Stability-HSI" and "Stability-Gravity".

Evaluation Datasets. We perform both quantitative and qualitative evaluations on our collected HSIBench dataset. To assess HSI simulation stability across different scenarios, we divide HSIBench into three levels of difficulty, *i.e*, easy, medium, and hard, based on interaction complexity.

4.1 Results and Analysis

Quantitative Evaluations. We first present quantitative comparisons of HSI reconstruction and simulation quality in Tab. 1. As shown, our method significantly outperforms the only existing baseline, HSfM, as well as the ablated variants, across all evaluated metrics. This demonstrates both the overall effectiveness of our approach and the contribution of the proposed components. Note that V1, V2, and V3 do not report scene penetration percentages, as their 3D reconstruction is identical to that of V1. We then report quantitative results on image-to-3D generation quality in Tab. 2. Importantly, our method (incorporating DSRO) achieves improved physical plausibility and interaction stability, along with superior geometric accuracy.

Qualitative Evaluations. **(1)** In Fig. 5, we present the qualitative comparisons with HSfM. Specifically, we apply Poisson reconstruction to the point maps generated by HSfM and place the reconstructed objects into the simulator for evaluation. As shown, HSfM often fails to produce stable human–object interactions: the human frequently kicks objects away and ends up standing alone. In contrast, our method consistently achieves stable interaction states within the simulation. **(2)** Fig. 4 further compares our approach with DSO and MIDI in

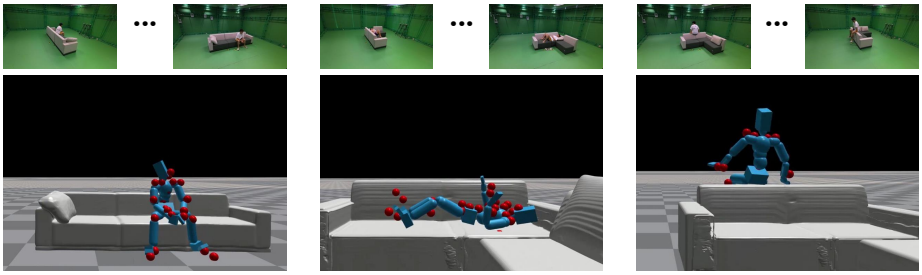

Fig. 6: Additional simulation results produced by HSImul3R. Our method faithfully reconstructs and handles complex interactions with the surrounding scene in the simulator.

terms of image-to-3D reconstruction quality. Both baselines struggle to recover accurate structures and often introduce geometric distortions, which in turn lead to instability during simulation. By contrast, our DSRO fine-tuned model mitigates these issues, yielding more structurally faithful and stable reconstructions. **(3)** In Fig. 6, we provide additional results reconstructed and simulated by our method. HSImul3R robustly handles intricate interactions with the surrounding scene in the simulator, with further examples provided in the supplementary materials and the demo video.

Real-world Robotics Deployment. Building upon the refined human motions detailed in Sec. 3.3, we utilize GMR [2, 92] to retarget human trajectories onto the Unitree G1 humanoid robots. These retargeted motions then serve as a prior for diffusion-guided reinforcement learning [35], which we employ to train a whole-body control policy within the IsaacGym simulator [49]. This framework allows the agent to learn robust balancing by leveraging the generative priors of diffusion models during the RL training phase. Once trained, the resulting control policy is deployed directly onto the physical G1 humanoid hardware via the Unitree SDK. As illustrated in Fig. 7, the successful deployment of the resulting policy on the physical Unitree G1 robot demonstrates that our refined motions facilitate robust robot-scene interactions. This framework provides a scalable foundation for leveraging vast, cost-effective datasets from platforms like YouTube to augment existing training data for large-scale embodied AI models. Ultimately, our approach promotes the execution of complex robotic tasks and supports various downstream applications in autonomous robot-scene interaction.

Analysis of Scene-targeted Simulation. Fig. 8 presents ablation studies on scene-targeted simulation loss defined in Eq. 5. Results indicate that removing the distance-minimization term destabilizes the humanoid, leading to exaggerated motions and often kicking objects away.

Analysis of the Number of Input Views. In Tab. 3, we analyze the effect of the number of input views. The results show that increasing the number of

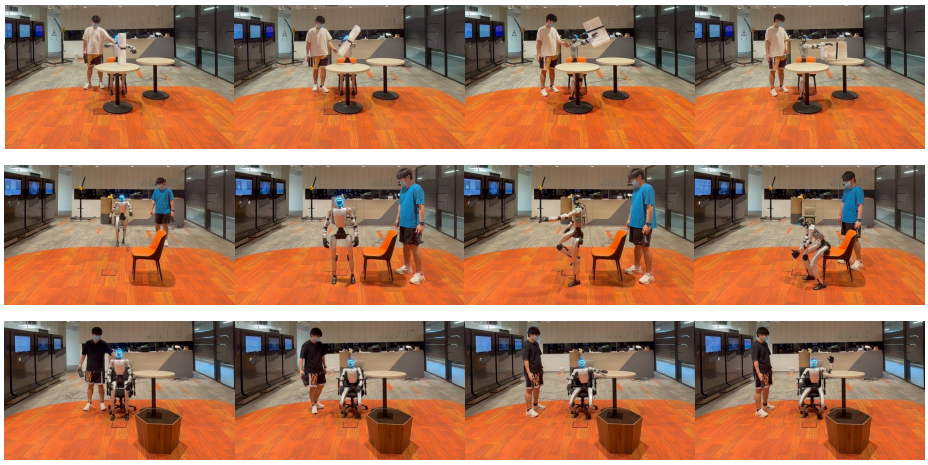

Fig. 7: Sim-to-Real results deployed on Unitree G1 humanoid robotics. The refined human motions by our method could be successfully transferred and deployed on a real-world humanoid robot to conduct various human-scene-interaction scenarios.

Table 3: Quantitative analysis of number of input views. We quantitatively evaluate reconstruction quality and simulation stability under different numbers of input views. While additional views slightly improve motion quality, they don't have obvious impact over the interaction stability and increase scene penetration.

Method	Stability-HSI (%) ↑			Scene Penetration	Human Motion Quality	
	Easy	Medium	Hard	SP-3D (%) ↓	W-MPJPE ↓	PA-MPJPE ↓
16-view	**55.16**	29.51	13.59	21.81	**4.01**	**1.99**
10-view	52.93	**32.17**	13.03	**21.00**	4.06	2.05
4-view	53.68	30.56	**13.92**	22.90	4.09	2.17

views yields slight improvements in human motion quality. Notably, however, the number of views has little impact on penetration handling or the overall stability of the simulation. This behavior is consistent with what our metrics capture: interaction stability, 3D human-scene penetration, and post-simulation motion quality are high-level 3D-quality measures rather than low-level quantities (*e.g.*, camera pose or depth) that scale with the number of views. Because HSImul3R refines contacts directly in both 3D space and the simulator, four views already provide sufficient information. This indicates that four views offer a favorable cost-performance trade-off.

Limitation and Failure Cases. While HSImul3R represents the first attempt at simulation-ready reconstruction of human–scene interactions, we acknowledge that our method has certain limitations: **(1)** the successful ratio is not very high, particularly in scenarios involving complex interactions or multiple objects

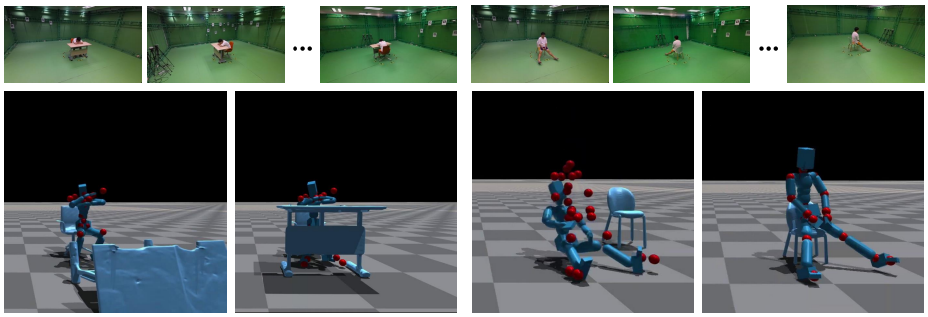

Fig. 8: Ablation studies on Eq. 5. Without the proposed scene-targeted RL, the simulation often results in unintended object displacement and fails to maintain stable interactions.

(more than three); **(2)** In many failure cases, the humanoid and objects tend to end up standing independently rather than engaging in meaningful interactions (see supplementary material); **(3)** Our fine-tuned image-to-3D model inevitably inherits biases from both the MIDI original training dataset and our collected HSIVBench, which may constrain its generalizability to out-of-domain cases.

5 Conclusion

In this work, we introduced HSImul3R, the first framework for reconstructing simulation-ready human–scene interactions from uncalibrated sparse views. Our approach incorporates a contact-aware interaction model to mitigate human–scene penetration issues in 3D reconstruction, a scene-targeted reinforcement learning strategy to promote stable interactions within the simulator, and a direct simulation reward optimization scheme that leverages simulation feedback to fine-tune the image-to-3D generative model, thereby improving simulation success rates. To support both training and evaluation, we collected the HSIBench dataset. Extensive experiments demonstrate that HSImul3R achieves high-fidelity results, delivering both stable simulations and high-quality image-to-3D reconstructions, and outperforms existing state-of-the-art methods.

Acknoledgements.. This study is supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOE-T2EP20223-0002). This research is also supported by cash and in-kind funding from NTU S-Lab and industry partner(s).

References

1. Allshire, A., Choi, H., Zhang, J., McAllister, D., Zhang, A., Kim, C.M., Darrell, T., Abbeel, P., Malik, J., Kanazawa, A.: Visual imitation enables contextual humanoid control. arXiv 2505.03729 (2025) 5, 9
2. Araujo, J.P., Ze, Y., Xu, P., Wu, J., Liu, C.K.: Retargeting matters: General motion retargeting for humanoid motion tracking. arXiv 2510.02252 (2025) 13
3. Barcellona, L., Zadaianchuk, A., Allegro, D., Papa, S., Ghidoni, S., Gavves, E.: Dream to manipulate: Compositional world models empowering robot imitation learning with imagination. In: ICLR (2025) 4
4. Bhatnagar, B.L., Xie, X., Petrov, I.A., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Behave: Dataset and method for tracking human object interactions. In: CVPR (2022) 2
5. Bhatnagar, B.L., Xie, X., Petrov, I.A., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: BEHAVE: dataset and method for tracking human object interactions. In: CVPR (2022) 2
6. Bian, H., Kong, L., Xie, H., Pan, L., Qiao, Y., Liu, Z.: DynamicCity: large-scale 4D occupancy generation from dynamic scenes. In: ICLR (2025) 4
7. Borycki, P., Smolak, W., Waczynska, J., Mazur, M., Tadeja, S.K., Spurek, P.: GASP: gaussian splatting for physic-based simulations. arXiv 2409.05819 (2024) 4
8. Cai, Z., Li, Z., Li, X., Li, B., Wang, Z., Zhang, Z., Xiu, Y.: Up2you: Fast reconstruction of yourself from unconstrained photo collections. arXiv 2509.24817 (2025) 2
9. Cai, Z., Yin, W., Zeng, A., Wei, C., Sun, Q., Yanjun, W., Pang, H.E., Mei, H., Zhang, M., Zhang, L., Loy, C.C., Yang, L., Liu, Z.: SMPLer-X: scaling up expressive human pose and shape estimation. In: NeurIPS (2023) 2
10. Cao, Y., Han, K., Wong, K.Y.K.: Sesdf: Self-evolved signed distance field for implicit 3d clothed human reconstruction. In: CVPR (2023) 2
11. Chen, Y., Chen, X., Xue, Y., Chen, A., Xiu, Y., Pons-Moll, G.: Human3r: Everyone everywhere all at once. arXiv 2510.06219 (2025) 2
12. Chen, Y., Xie, T., Zong, Z., Li, X., Gao, F., Yang, Y., Wu, Y.N., Jiang, C.: Atlas3D: physically constrained self-supporting text-to-3d for simulation and fabrication. In: NeurIPS (2024) 4
13. Chen, Z., Moon, G., Guo, K., Cao, C., Pidhorskyi, S., Simon, T., Joshi, R., Dong, Y., Xu, Y., Pires, B., Wen, H., Evans, L., Peng, B., Buffalini, J., Trimble, A., McPhail, K., Schoeller, M., Yu, S., Romero, J., Zollhöfer, M., Sheikh, Y., Liu, Z., Saito, S.: Urhand: Universal relightable hands. In: CVPR (2024) 2
14. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T.A., Nießner, M.: ScanNet: richly-annotated 3D reconstructions of indoor scenes. In: CVPR (2017) 4
15. Fan, Z., Parelli, M., Kadoglou, M.E., Chen, X., Kocabas, M., Black, M.J., Hilliges, O.: HOLD: category-agnostic 3D reconstruction of interacting hands and objects from video. In: CVPR (2024) 2
16. Fu, Z., Zhao, Q., Wu, Q., Wetzstein, G., Finn, C.: HumanPlus: humanoid shadowing and imitation from humans. In: CoRL (2024) 4
17. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., Malik, J.: Humans in 4D: Reconstructing and tracking humans with transformers. In: ICCV (2023) 6
18. He, T., Gao, J., Xiao, W., Zhang, Y., Wang, Z., Wang, J., Luo, Z., He, G., Sobanbab, N., Pan, C., Yi, Z., Qu, G., Kitani, K., Hodgins, J.K., Fan, L., Zhu, Y., Liu, C., Shi, G.: ASAP: aligning simulation and real-world physics for learning agile humanoid whole-body skills. In: RSS (2025) 5

19. He, T., Xiao, W., Lin, T., Luo, Z., Xu, Z., Jiang, Z., Kautz, J., Liu, C., Shi, G., Wang, X., Fan, L., Zhu, Y.: HOVER: versatile neural whole-body controller for humanoid robots. arXiv 2410.21229 (2024) 4
20. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: low-rank adaptation of large language models. In: ICLR (2022) 10
21. Huang, B., Yu, Z., Chen, A., Geiger, A., Gao, S.: 2d gaussian splatting for geometrically accurate radiance fields. In: SIGGRAPH (2024) 9
22. Huang, P., Matzen, K., Kopf, J., Ahuja, N., Huang, J.: DeepMVS: learning multi-view stereopsis. In: CVPR (2018) 4
23. Huang, Z., Guo, Y., An, X., Yang, Y., Li, Y., Zou, Z., Liang, D., Liu, X., Cao, Y., Sheng, L.: MIDI: multi-instance diffusion for single image to 3D scene generation. In: CVPR (2025) 5, 7, 10, 11
24. Jiang, N., Liu, T., Cao, Z., Cui, J., Zhang, Z., Chen, Y., Wang, H., Zhu, Y., Huang, S.: Full-body articulated human-object interaction. In: ICCV (2023) 2
25. Jiang, N., Zhang, Z., Li, H., Ma, X., Wang, Z., Chen, Y., Liu, T., Zhu, Y., Huang, S.: Scaling up dynamic human-scene interaction modeling. In: CVPR (2024) 2
26. Jiang, Y., Yu, C., Xie, T., Li, X., Feng, Y., Wang, H., Li, M., Lau, H.Y.K., Gao, F., Yang, Y., Jiang, C.: VR-GS: A physical dynamics-aware interactive gaussian splatting system in virtual reality. In: SIGGRAPH (2024) 4
27. Kaneko, T.: Improving physics-augmented continuum neural radiance field-based geometry-agnostic system identification with lagrangian particle optimization. In: CVPR (2024) 4
28. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3D gaussian splatting for real-time radiance field rendering. ACM TOG **42**(4), 139:1–139:14 (2023) 2, 4
29. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: ICCV (2023) 7
30. Li, J., Bian, S., Xu, C., Chen, Z., Yang, L., Lu, C.: HybrIK-X: hybrid analytical-neural inverse kinematics for whole-body mesh recovery. IEEE TPAMI **47**(4), 2754–2769 (2025) 2
31. Li, R., Zheng, C., Rupprecht, C., Vedaldi, A.: DSO: aligning 3D generators with simulation feedback for physical soundness. In: ICCV (2025) 4, 8, 10, 11
32. Li, X., Qiao, Y., Chen, P.Y., Jatavallabhula, K.M., Lin, M.C., Jiang, C., Gan, C.: PAC-NeRF: physics augmented continuum neural radiance fields for geometry-agnostic system identification. In: ICLR (2023) 4
33. Li, Y., Jiang, L., Xu, L., Xiangli, Y., Wang, Z., Lin, D., Dai, B.: MatrixCity: A large-scale city dataset for city-scale neural rendering and beyond. In: ICCV (2023) 4
34. Li, Z., Tucker, R., Cole, F., Wang, Q., Jin, L., Ye, V., Kanazawa, A., Holynski, A., Snavely, N.: Megasam: Accurate, fast and robust structure and motion from casual dynamic videos. In: CVPR (2025) 9
35. Liao, Q., Truong, T.E., Huang, X., Tevet, G., Sreenath, K., Liu, C.K.: Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion. arXiv 2508.08241 (2025) 5, 13
36. Liu, J., Cao, Y., Yang, T., Xu, Z., Keppo, J., Shan, Y., Qie, X., Shou, M.Z.: HOSNeRF: dynamic human-object-scene neural radiance fields from a single video. In: ICCV (2023) 2
37. Liu, X., Xie, H., Zhang, S., Yao, H., Ji, R., Nie, L., Tao, D.: 2D semantic-guided semantic scene completion. IJCV **133**(3), 1306–1325 (2025) 4
38. Liu, X., Liang, J., Lin, Z., Hou, H., Li, Y., Lu, C.: ImDy: human inverse dynamics from imitated observations. In: ICLR (2025) 5

39. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv 1711.05101 (2017) 10
40. Lu, J., Huang, C.H.P., Bhattacharya, U., Huang, Q., Zhou, Y.: Humoto: A 4d dataset of mocap human object interactions. In: ICCV (2025) 2
41. Lu, J., Huang, C.P., Bhattacharya, U., Huang, Q., Zhou, Y.: HUMOTO: A 4D dataset of mocap human object interactions. In: ICCV (2025) 2
42. Luo, R., Geng, H., Deng, C., Li, P., Wang, Z., Jia, B., Guibas, L.J., Huang, S.: PhysPart: physically plausible part completion for interactable objects. arXiv 2408.13724 (2024) 4
43. Luo, Z., Cao, J., Christen, S., Winkler, A., Kitani, K., Xu, W.: Omnigrasp: grasping diverse objects with simulated humanoids. In: NeurIPS (2024) 4
44. Luo, Z., Cao, J., Merel, J., Winkler, A., Huang, J., Kitani, K.M., Xu, W.: Universal humanoid motion representations for physics-based control. In: ICLR (2024) 4
45. Luo, Z., Cao, J., Winkler, A., Kitani, K., Xu, W.: Perpetual humanoid control for real-time simulated avatars. In: ICCV (2023) 4, 8, 9, 10
46. Luo, Z., Wang, J., Liu, K., Zhang, H., Tessler, C., Wang, J., Yuan, Y., Cao, J., Lin, Z., Wang, F., Hodgins, J.K., Kitani, K.: SMPLOlympics: sports environments for physically simulated humanoids. arXiv 2407.00187 (2024) 5
47. Luo, Z., Wang, J., Liu, K., Zhang, H., Tessler, C., Wang, J., Yuan, Y., Cao, J., Lin, Z., Wang, F., Hodgins, J.K., Kitani, K.: SMPLOlympics: sports environments for physically simulated humanoids. arXiv 2407.00187 (2024) 5
48. Makoviychuk, V., Wawrzyniak, L., Guo, Y., Lu, M., Storey, K., Macklin, M., Hoeller, D., Rudin, N., Allshire, A., Handa, A., State, G.: Isaac Gym: High performance GPU based physics simulation for robot learning. In: NeurIPS (2021) 8
49. Makoviychuk, V., Wawrzyniak, L., Guo, Y., Lu, M., Storey, K., Macklin, M., Hoeller, D., Rudin, N., Allshire, A., Handa, A., et al.: Isaac gym: High performance gpu-based physics simulation for robot learning. arXiv 2108.10470 (2021) 13
50. Mezghanni, M., Bodrito, T., Boulkenafed, M., Ovsjanikov, M.: Physical simulation layer for accurate 3D modeling. In: CVPR (2022) 4
51. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: representing scenes as neural radiance fields for view synthesis. In: ECCV (2020) 2
52. Müller, L., Choi, H., Zhang, A., Yi, B., Malik, J., Kanazawa, A.: Reconstructing people, places, and cameras. In: CVPR (2025) 2, 6, 10, 12
53. Ni, J., Chen, Y., Jing, B., Jiang, N., Wang, B., Dai, B., Li, P., Zhu, Y., Zhu, S., Huang, S.: PhyRecon: physically plausible neural scene reconstruction. In: NeurIPS (2024) 4
54. Nie, Y., Han, X., Guo, S., Zheng, Y., Chang, J., Zhang, J.: Total3DUnderstanding: joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In: CVPR (2020) 4
55. Pan, L., Yang, Z., Dou, Z., Wang, W., Huang, B., Dai, B., Komura, T., Wang, J.: TokenHSI: unified synthesis of physical human-scene interactions through task tokenization. In: CVPR (2025) 2
56. Park, J.J., Florence, P.R., Straub, J., Newcombe, R.A., Lovegrove, S.: DeepSDF: learning continuous signed distance functions for shape representation. In: CVPR (2019) 4
57. Pavlakos, G., Shan, D., Radosavovic, I., Kanazawa, A., Fouhey, D., Malik, J.: Reconstructing hands in 3d with transformers. In: CVPR (2024) 2

58. Peng, X.B., Abbeel, P., Levine, S., van de Panne, M.: DeepMimic: example-guided deep reinforcement learning of physics-based character skills. ACM TOG **37**(4), 143 (2018) 4
59. Peng, X.B., Kanazawa, A., Malik, J., Abbeel, P., Levine, S.: Sfv: Reinforcement learning of physical skills from videos. ACM TOG **37**(6), 1–14 (2018) 4
60. Peng, X.B., Ma, Z., Abbeel, P., Levine, S., Kanazawa, A.: AMP: adversarial motion priors for stylized physics-based character control. ACM TOG **40**(4), 144:1–144:20 (2021) 4
61. Pun, A., Deng, K., Liu, R., Ramanan, D., Liu, C., Zhu, J.: Generating physically stable and buildable brick structures from text. In: ICCV (2025) 4
62. Ravi, N., Gabeur, V., Hu, Y., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C., Girshick, R.B., Dollár, P., Feichtenhofer, C.: SAM 2: Segment anything in images and videos. In: ICLR (2025) 6, 9
63. Ren, J., Yu, C., Chen, S., Ma, X., Pan, L., Liu, Z.: DiffMimic: efficient motion mimicking with differentiable physics. In: ICLR (2023) 4
64. Ren, J., Zhang, M., Yu, C., Ma, X., Pan, L., Liu, Z.: InsActor: instruction-driven physics-based characters. In: NeurIPS (2023) 4
65. Schönberger, J.L., Frahm, J.: Structure-from-motion revisited. In: CVPR (2016) 4
66. Seitz, S.M., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: CVPR (2006) 4
67. Siyao, L., Feng, Y., Tehari, O., Loy, C.C., Black, M.J.: Half-Physics: enabling kinematic 3D human model with physical interactions. arXiv 2507.23778 (2025) 5
68. Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.A.: Semantic scene completion from a single depth image. In: CVPR (2017) 4
69. Tevet, G., Raab, S., Cohan, S., Reda, D., Luo, Z., Peng, X.B., Bermano, A.H., van de Panne, M.: CLoSD: closing the loop between simulation and diffusion for multi-task character control. In: ICLR (2025) 4, 11
70. Tian, Y., Zhang, H., Liu, Y., Wang, L.: Recovering 3D human mesh from monocular images: A survey. IEEE TPAMI **45**(12), 15406–15425 (2023) 2
71. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotný, D.: VGGT: visual geometry grounded transformer. In: CVPR (2025) 4
72. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: DUSt3R: geometric 3D vision made easy. In: CVPR (2024) 2, 4, 6
73. Wang, Y., Lin, J., Zeng, A., Luo, Z., Zhang, J., Zhang, L.: PhysHOI: physics-based imitation of dynamic human-object interaction. arXiv 2312.04393 (2023) 4
74. Wang, Y., Wang, Z., Liu, L., Daniilidis, K.: Tram: Global trajectory and motion of 3d humans from in-the-wild videos. In: ECCV. pp. 467–487. Springer (2024) 9
75. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3D latents for scalable and versatile 3D generation. In: CVPR (2025) 11
76. Xie, H., Chen, Z., Hong, F., Liu, Z.: CityDreamer: compositional generative model of unbounded 3D cities. In: CVPR (2024) 4
77. Xie, H., Chen, Z., Hong, F., Liu, Z.: Generative gaussian splatting for unbounded 3D city generation. In: CVPR (2025) 4
78. Xie, H., Chen, Z., Hong, F., Liu, Z.: Compositional generative model of unbounded 4D cities. IEEE TPAMI **48**(1), 312–328 (2026) 4
79. Xie, H., Wen, B., Zheng, J., Chen, Z., Hong, F., Diao, H., Liu, Z.: DynamicVLA: A vision-language-action model for dynamic object manipulation. arXiv 2601.22153 (2026) 2

80. Xie, H., Yao, H., Zhou, S., Mao, J., Zhang, S., Sun, W.: GRNet: gridding residual network for dense point cloud completion. In: ECCV (2020) 4
81. Xie, T., Zong, Z., Qiu, Y., Li, X., Feng, Y., Yang, Y., Jiang, C.: PhysGaussian: physics-integrated 3D gaussians for generative dynamics. In: CVPR (2024) 4
82. Xie, X., Wen, B., Chang, Y., Rabeti, H., Li, J., Yuan, Y., Pons-Moll, G., Birchfield, S.: Cari4d: Category agnostic 4d reconstruction of human-object interaction. arXiv 2512.11988 (2025) 2
83. Xiu, Y., Yang, J., Cao, X., Tzionas, D., Black, M.J.: Econ: Explicit clothed humans optimized via normal integration. In: CVPR (2023) 2
84. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: ViTPose: simple vision transformer baselines for human pose estimation. In: NeurIPS (2022) 6
85. Yan, H., Zhang, M., Li, Y., Ma, C., Ji, P.: PhyCAGE: physically plausible compositional 3D asset generation from a single image. arXiv 2411.18548 (2024) 4
86. Yan, Y., Lin, H., Zhou, C., Wang, W., Sun, H., Zhan, K., Lang, X., Zhou, X., Peng, S.: Street Gaussians: Modeling dynamic urban scenes with gaussian splatting. In: ECCV (2024) 4
87. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth Anything: Unleashing the power of large-scale unlabeled data. In: CVPR (2024) 4
88. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything V2. In: NeurIPS (2024) 4
89. Yang, L., Li, K., Zhan, X., Lv, J., Xu, W., Li, J., Lu, C.: ArtiBoost: boosting articulated 3D hand-object pose estimation via online exploration and synthesis. In: CVPR (2022) 2
90. Yang, Y., Jia, B., Zhi, P., Huang, S.: PhyScene: physically interactable 3D scene synthesis for embodied AI. In: CVPR (2024) 4
91. Yin, S., Ze, Y., Yu, H.X., Liu, C.K., Wu, J.: Visualmimic: Visual humanoid loco-manipulation via motion tracking and generation. arXiv 2509.20322 (2025) 2
92. Ze, Y., Araújo, J.P., Wu, J., Liu, C.K.: Gmr: General motion retargeting (2025), https://github.com/YanjieZe/GMR 2, 13
93. Ze, Y., Chen, Z., Araújo, J.P., Cao, Z., Peng, X.B., Wu, J., Liu, C.K.: TWIST: teleoperated whole-body imitation system. arXiv 2505.02833 (2025) 4
94. Ze, Y., Chen, Z., Araújo, J.P., ang Cao, Z., Peng, X.B., Wu, J., Liu, C.K.: Twist: Teleoperated whole-body imitation system. arXiv2505.02833 (2025) 2