SLoHa: A 4D Human Locomotion Dataset of Scene-constrained Large-object Handling via a Democratic Data Acquisition Solution

Anonymous authors
Paper under double-blind review

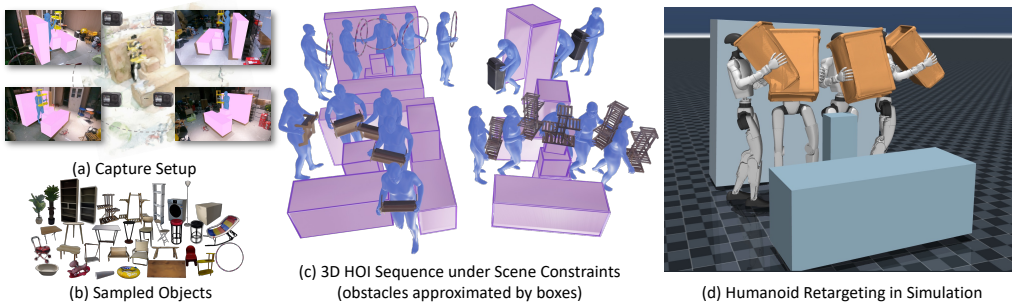


Figure 1: We introduce SLoHa, the first 4D human–scene–object interaction dataset for scene-constrained large-object handling, collected with a low-cost multi-view acquisition solution. It provides reconstructed human motion (SMPLX), scene constraints(approximated by boxes), object geometry and 6D pose sequence, covering 152 everyday large objects and diverse obstacle layouts. The collected interactions can further serve as motion references for humanoid simulation.

Abstract

Carrying large objects through cluttered environments is an important capability for robots in daily assistance, logistics, and workplaces, requiring coordinated whole-body motion, object control, and obstacle avoidance. Yet learning this capability remains underexplored, partly due to the scarcity of suitable training data. Human demonstrations provide a valuable reference for robots to learn such coordinated behaviors. However, existing datasets primarily capture small-object interactions in unconstrained environments and often rely on expensive motion-capture systems that are difficult to scale, limiting the availability of scene-aware large-object demonstrations and hindering data-driven learning of this capability. To bridge this gap, we present SLoHa, a dataset for scene-constrained large-object handling, together with a democratic data acquisition solution that makes such demonstrations more accessible to collect. Using four exocentric cameras, modern reconstruction priors, and sparse human-assisted alignment, our pipeline recovers human motion, object geometry and 6D pose trajectories, and scene constraints in a shared coordinate system without specialized MoCap suits or object markers. SLoHa contains 21.4 hours of demonstrations involving 152 large objects and 435 obstacle layouts. We present a scene-conditioned large object handling benchmark. Experiments show that our dataset improves constraint-compliant motion generation, validating its effectiveness for studying scene-aware interaction and its potential use in embodied applications.

1 Introduction

Enabling robots to interact with objects in complex environments is essential for their deployment in daily assistance, logistics, and workplaces. Among these interactions, handling large objects (e.g., bulky boxes, large containers, and furniture) is a pivotal capability that requires coordinated whole-body motion, object control, and adaptation to surrounding obstacles. In everyday settings,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

transporting large objects often involves navigating narrow passages, turning corners, and avoiding surrounding obstacles. These applications motivate the study of coordinated human-object motion under scene constraints. 4D human-object interaction (HOI) data captures how people coordinate body motion and object orientation to navigate surrounding geometry, providing demonstrations for robots to learn scene-constrained large-object handling.

However, existing 4D HOI datasets predominantly focus on small-object manipulation, such as fine-grained grasping or tool usage (Taheri et al., 2020; Jiang et al., 2024b;a). While recent works (Lu et al., 2025; Liu et al., 2025; Li et al., 2023) have begun to incorporate large objects, they are restricted to unconstrained environments with only a small number of large instances. The limited coverage of large-object interactions under scene constraints leaves a significant data gap for scene-aware HOI modeling and downstream robotic applications.

To fill this void, it is imperative to construct an HOI dataset dedicated to scene-constrained large-object handling. However, scaling up such data remains challenging. Prevailing methodologies rely heavily on specialized motion capture (MoCap) systems (Lu et al., 2025; Liu et al., 2025; Jiang et al., 2024a). Although capable of capturing high-quality human motion sequences, MoCap systems necessitate synchronized sensors and meticulous subject/object instrumentation. This results in high operational costs and spatial confinement, precluding data acquisition in those natural, cluttered environments that are essential for learning scene-constrained interaction strategies. Building on these observations, our goal is to capture human interactions with large objects under scene constraints using a low-cost and easily deployable capture system.

To democratize data acquisition, we advocate for a deliberate trade-off: minimizing hardware complexity and human effort while leveraging modern algorithmic priors. Effectively, we shift the burden from expensive hardware and labor-intensive processes to advanced algorithms. Specifically, we leverage recent breakthroughs in 3D reconstruction, such as VGGT (Wang et al., 2025) and ReconViaGen (Chang et al., 2025) for feed-forward multi-view reconstruction, FoundationPose (Wen et al., 2024) for model-based tracking, and SMPLX (Pavlakos et al., 2019) for full-body human fitting. Additionally, we incorporate a human-in-the-loop alignment module that uses sparse annotations to correct visible pose misalignment; the proportion of sequences requiring manual annotation and the alignment annotation time are reported as part of the acquisition cost. This allows us to recover globally coherent 3D *human-object-scene* information from a lightweight, sparse multi-view setup, entirely eliminating the need for specialized MoCap hardware, object markers (Liu et al., 2025), or manual CAD modeling (Lu et al., 2025). Through this democratic solution, we reconstruct the human body, object geometry, and static scene into a coherent coordinate system. In total, our pipeline has enabled the scalable collection of 21.4 hours of human demonstrations in structured real environments, encompassing 435 corridor-like pathway obstacle layouts and 152 everyday large objects spanning diverse categories, shapes, and volumes.

Another critical challenge lies in efficiently scaling scene diversity. To address this, we use cardboard boxes as geometric proxies for real-world obstacles. Although such a simplification abstracts away fine-grained object geometry, it preserves the essential volumetric constraints that dictate large-object transportation. Besides, SLoHa follows a capture-based route: the obstacles are physically present during capture, so the demonstrations record how people adapt their posture, footsteps, and carried-object orientation to spatial constraints. Our emphasis is therefore on capturing scene-induced motion, rather than scene geometry alone.

To assess the utility of the collected data, we organize the evaluation around a central question: *can SLoHa support learning scene-aware large-object handling?* First, we establish a scene-conditioned HOI generation benchmark by augmenting a diffusion-based baseline with explicit scene information from SLoHa. Comparison with an otherwise matched no-scene baseline isolates the contribution of scene conditioning to adherence to environmental constraints. Second, we evaluate human-object motion forecasting on unseen large objects, comparing the same backbone trained on different datasets to assess the generalization benefits of training on SLoHa. Finally, as a complementary robot-level evaluation, we retarget human-object interaction sequences from SLoHa to a humanoid robot and evaluate their execution in physics simulation. We additionally present real-world demonstrations on a Unitree G1 humanoid.

In summary, our main contributions are:

Table 1: Comparison of our dataset with existing human-object and human-scene interaction datasets.

Name	Hours	Specialized MoCap Hardware	Scene Constraint	Dynamic Object	#Scene	#Obj
PROX (Hassan et al., 2019)	0.9	×	✓	×	12	-
GRAB (Taheri et al., 2020)	3.8	✓	×	✓	-	51
BEHAVE (Bhatnagar et al., 2022)	4.2	×	×	✓	-	20
NeuralDome (Zhang et al., 2023)	4.6	✓	×	✓	-	23
Circle (Araújo et al., 2023)	10.0	✓	✓	×	9	-
OMOMO (Li et al., 2023)	10.0	✓	×	✓	-	15
TRUMANS (Jiang et al., 2024b)	15.0	✓	✓	✓	100	20
LINGO (Jiang et al., 2024a)	16.0	✓	✓	✓	120	20
ParaHome (Kim et al., 2025)	8.1	✓	✓	✓	1	22
Core4D-Real (Liu et al., 2025)	2.8	✓	×	✓	-	37
HUMOTO (Lu et al., 2025)	2.2	✓	×	✓	-	63
Ours	**21.4**	×	✓	✓	**435**	**152**

- We construct a new 4D human-scene-object interaction dataset that is, to the best of our knowledge, the first dataset focusing on large-object handling under real-world scene constraints.
- We propose a democratic solution based on a lightweight real-capture setup and algorithmic data processing, that supports scalable, parallel acquisition of 4D human-object-scene interaction data.
- We conduct comprehensive experiments on scene-aware large-object interaction synthesis and physics-based humanoid retargeting for HOI simulation, empirically validating our dataset's ability to support these complex tasks.
- Our dataset, pipeline, and algorithm will be publicly available.

2 Related Work

Human Motion Datasets. The early exploration of human motion modeling has contributed to the development of several high-resolution optical motion capture datasets such as CMU (Mocap, 2003) and Human3.6M (Ionescu et al., 2013). However, the scale of these datasets remains limited due to the complexity and high cost of the data acquisition systems. To address this issue, AMASS (Mahmood et al., 2019) unified multiple previous motion capture datasets based on optimization algorithms, which has inspired the scaling up of later human motion data collection (Cai et al., 2022; Xiong et al., 2024). As generative models (Kingma & Welling, 2013; Ho et al., 2020) witness rapid progress, researchers are delving into human motion synthesis. Text-to-motion is one of the most attractive topics in conditional motion generation, where motions are generated based on textual descriptions. The text-aligned motion generation task benefits from the BABEL (Punnakkal et al., 2021) and HumanML3D (Guo et al., 2022) datasets, which contribute language labels to the existing marker-based mocap data through crowdsourced data collection. Follow-up works Motion-X (Lin et al., 2023) and Go-to-Zero (Fan et al., 2025) collect human motion data from web-scale video sources, and to facilitate dataset scaling, these works leverage pretrained SMPLX estimation models and optimize global motion to guarantee dataset quality. Additionally, HUMANISE (Wang et al., 2022) and TeSMo (Yi et al., 2024) collect human motion data in various 3D scenes with text annotations for text-controlled, scene-aware motion generation. These works align the captured human motion sequences from AMASS and HumanML3D with the existing scanned (Dai et al., 2017) or synthetic (Fu et al., 2021a) indoor scenes. However, these works only consider human activities within static surroundings, thus overlooking broader interactive engagements with dynamic objects.

Human-object Interaction Datasets. This line of work collects both human motion and object motion. GRAB (Taheri et al., 2020) is the first datasets that addresses the full-body human-object interaction. BEHAVE (Bhatnagar et al., 2022) and OMOMO (Li et al., 2023) present the complex daily human-object contact scenarios and record interactions with larger-sized objects. However, the limited data for each object impedes the model's capacity to generalize to unseen object categories. Subsequent works, such as IMHD^2 (Zhao et al., 2024), focus specifically on highly dynamic human-object interactions involving sports activities. ParaHome (Kim et al., 2025) designs a sophis-



Figure 2: **Capture setup.** Four exocentric GoPro cameras record large-object handling in modular obstacle layouts. The cameras cost approximately $1500 USD in total.

ticated home capture system with 3D ArUco markers to capture dynamic 3D movements of humans and objects within a typical home environment. HUMOTO (Lu et al., 2025) captures human-object interaction sequences involving multiple objects, with a specific focus on detailed hand poses. More recent efforts target the more complex scenario of human-object-human interaction in real-world settings. Core4D (Liu et al., 2025) and InterVLA (Xu et al., 2025) both use a hybrid RGB-Mocap system composed of two egocentric RGB cameras, four or five exocentric RGB cameras, and an OptiTrack MoCap system. Despite these hybrid designs, current frameworks still rely on costly optical motion capture components, alongside additional industrial reconstruction or artist-crafted object geometries. This reliance impedes the scalability of human-object interaction datasets. Moreover, these works also fail to take scene constraints into account in their data collection process. We argue that modeling large object manipulation necessitates considering scene constraints. In this work, we aim to propose a low-cost solution for capturing human-object interactions to expand dataset scale while incorporating scene constraints.

Human-scene-object interaction. Humans naturally interact with both objects and surrounding environments, engaging in various social activities. However, the acquisition and alignment of human, object, and corresponding scene geometries are highly challenging, which results in a scarcity of relevant data. Existing human–scene–object datasets attempt to fill this gap but still suffer from inherent limitations. HOI-M3 (Zhang et al., 2024) dataset captures human–object interactions within a contextual environment, with a focus on scenarios involving multiple humans and objects. However, this work treats the scene as a set of discrete objects rather than an integrated environment. TRUMANS (Jiang et al., 2024b) and LINGO (Jiang et al., 2024a) explore the collection of human–object interaction data integrated with indoor scene geometry via a VICON motion capture system. However, the object inventory within these datasets is restricted to small-scale items. SLoHa focuses on single-person large-object transportation through modular passage constraints, with reconstructed human motion, object geometry and trajectories, and obstacle geometry in a shared coordinate system. Our contribution lies in the task-focused coverage and acquisition pipeline of SLoHa, complementing these existing resources.

3 Dataset Construction

SLoHa captures scene-constrained large-object handling with an accessible multi-view system; Table 1 compares its coverage with existing datasets.

3.1 Data Capture

The capture design combines inexpensive cameras with modular physical obstacles. Four static viewpoints provide complementary observations of the subject and the carried object, which may occlude one another during transportation. Rather than prescribing a detailed motion trajectory, we vary the object and the surrounding passage so that the demonstrations contain human responses to different spatial constraints. The same acquisition procedure can be deployed in separate rooms without instrumenting the subjects or attaching markers to each object.

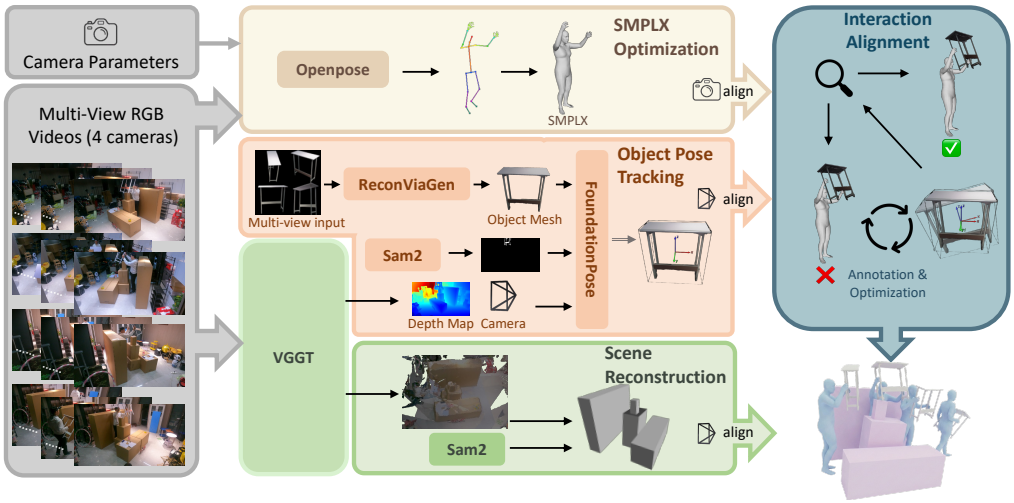


Figure 3: **Processing pipeline.** Multi-view videos yield aligned scene geometry, human motion, and object poses, followed by human-assisted interaction alignment.

We record synchronized 4K videos at 30 FPS over an area of approximately 10m^2 (Fig. 2). Reconstruction uses four exocentric GoPro cameras; an additional head-mounted view is recorded but not used. Subjects naturally transport designated objects through obstacle layouts under task-level instructions specifying the goal and collision-avoidance requirements. We recruit two subjects per room configuration, provide approximately one day of familiarization, and collect data over roughly three months using two systems in two rooms. Collection and calibration details are provided in Appendix A.1.

3.2 4D Data Reconstruction

Our pipeline (Fig. 3) reconstructs the scene, SMPL-X human motion, and object geometry and 6D trajectories in a shared coordinate system.

The output separates static surroundings from the two moving components of the interaction. Scene geometry describes the available space, the human representation describes whole-body locomotion, and the object mesh and pose sequence describe the volume swept by the carried object. Aligning these components is necessary because a plausible human trajectory alone does not determine whether a large object can pass through the same space. This representation supports joint human–object modeling and the scene-penetration evaluations below.

Scene reconstruction. VGGT (Wang et al., 2025) provides camera poses, depth maps, and point maps. We represent obstacles with 3D bounding boxes that describe occupancy and passage clearance. Avoiding an enclosing box also avoids the enclosed geometry, provided both are accurately aligned; this geometric property does not alone guarantee dynamic feasibility.

The bounding-box representation matches the modular capture setup and retains the dimensions most directly relevant to transportation: corridor width, obstacle height, and the clearance available for turning or lifting an object. It also gives different layouts a consistent representation for downstream conditioning. This abstraction omits fine surface structure and is not intended to model detailed contact with the obstacles. Our emphasis is on the coordinated motion needed to navigate around their occupied volumes.

SMPLX optimization. Following Xiong et al. (2024); Li et al. (2025), we triangulate multi-view OpenPose keypoints (Cao et al., 2017), fit per-frame SMPL-X body pose and shape using EasyMocap (eas, 2021) under multi-view keypoint constraints, and refine poses with a learned body prior to suppress implausible joint configurations.

The calibrated views constrain the same body estimate from complementary directions, while the parametric representation provides a common format across subjects and sequences. We retain the fitted whole-body motion alongside the independently estimated object trajectory, allowing downstream models to study their coordination rather than infer object motion from the body alone.

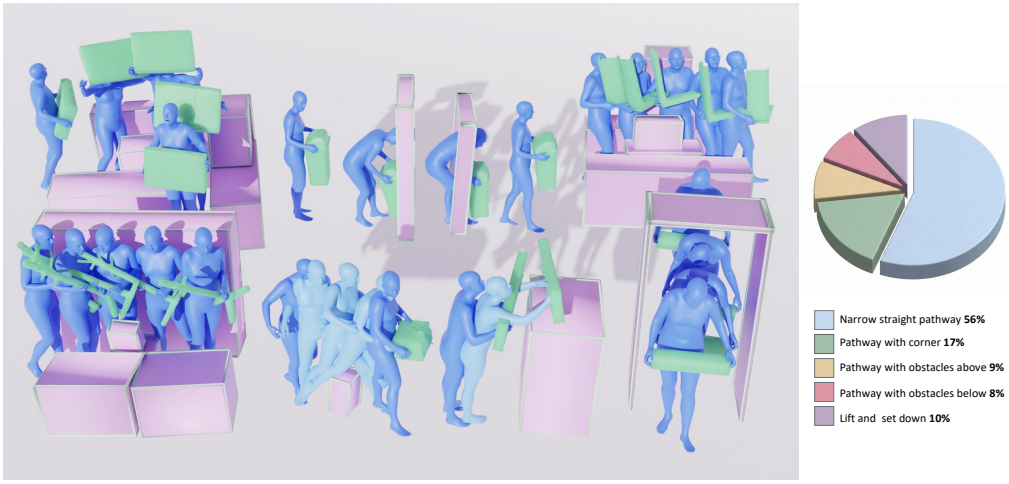


Figure 4: **Dataset gallery and distribution.** Left: interactions across the five scene-layout types. Right: their distribution in SLoHa.

Object pose tracking. ReconViaGen (Chang et al., 2025) reconstructs reusable object meshes from surrounding videos. SAM2 (Ravi et al., 2024) tracks object masks across frames. We align the reconstructed mesh with the scene-derived object point cloud to resolve their relative scale, then use FoundationPose (Wen et al., 2024) to estimate 6D poses from the mesh, masks, depth maps, and camera parameters. Object-scale alignment is described in Appendix A.1.

Reconstructing each object from a short surrounding video avoids requiring a manually built CAD model for every new instance. The resulting mesh can be reused across recordings involving the same object. Mask tracking maintains the association with that object over the interaction sequence, and the estimated rotation and translation place its geometry in the shared scene frame. Since reconstruction and tracking can still introduce visible misalignment, we subsequently apply the lightweight alignment procedure below rather than assume that the independent estimates are always mutually consistent.

3.3 Interaction Alignment

We correct visible object-pose misalignment in two stages. First, sparse 2D correspondences across camera views initialize the object pose in 3D. Second, hand–object correspondences refine the pose while regularization keeps it close to the network prediction. This human-assisted procedure improves geometric consistency without dense annotation. The annotation procedure and optimization objective are detailed in Appendix A.2.

Annotators mark corresponding object points in a reference frame rather than edit a complete 3D trajectory. The refined reference pose initializes subsequent frames; the second stage balances interaction consistency against the original pose estimate. This keeps human input sparse while retaining an explicit mechanism for correcting visible reconstruction errors.

3.4 Statistics

Object categories. SLoHa contains 152 objects across seven functional categories: Sports/Toys, Storage, Tables, Seating/Bedding, Household Tools, Racks/Carts, and Decor (Fig. 5). Selected for two-handed carrying, their elongated, wide, tall, and irregular shapes encourage varied grasp placement, torso stabilization, and motion strategies.

Scene layouts. The 435 layouts cover straight passages, corners, overhead obstacles, ground-level obstacles, and lift-and-set-down scenarios (Fig. 4), eliciting diverse whole-body and object-motion adaptations. These variations complement object diversity: the same transport goal may require different strategies depending on both the object's extent and the surrounding free space.

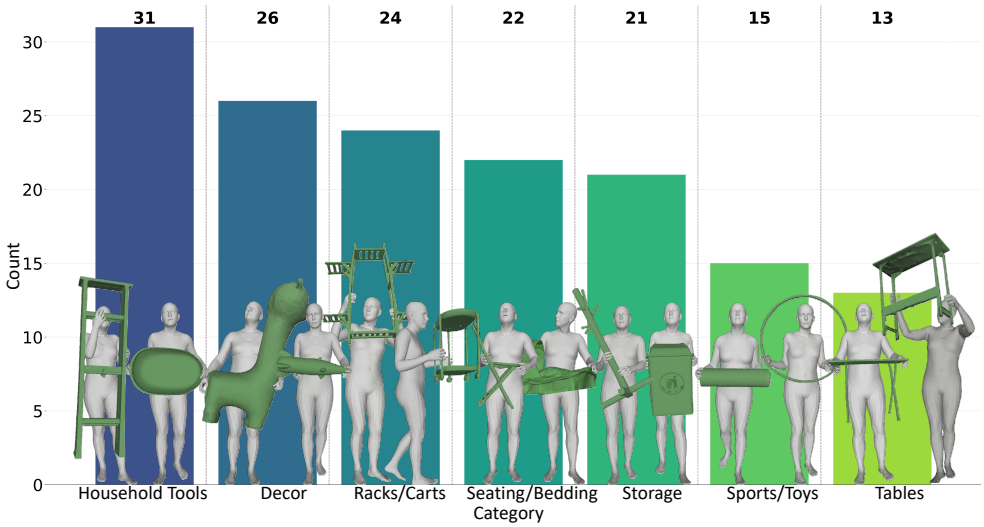


Figure 5: **Object distribution and examples.** The 152 large objects span seven functional categories and diverse geometries.

Table 2: Motion-quality comparison.

Dataset	Human motion					Object motion
	Foot sliding (cm) ↓	Jerk (m/s^3) ↓	MSNR (dB)	Coherence ↑	Diversity ↑	Jerk (m/s^3) ↓
OMOMO(Li et al., 2023)	2.130	15.10	12.37	0.619	**0.978**	27.40
HUMOTO(Lu et al., 2025)	0.958	**1.87**	9.42	**0.653**	0.956	1.13
SLoHa(Ours)	**0.658**	4.33	21.92	0.522	0.971	**0.76**

Motion Quality. Table 2 reports motion metrics following Lu et al. (2025). SLoHa achieves the lowest foot sliding and object jerk among the compared datasets, while retaining a diversity score close to OMOMO. Its higher human-motion jerk than HUMOTO and lower coherence may partly reflect the heterogeneous, scene-induced adjustments required for large-object handling, such as turning, bending, and stepping around obstacles, rather than reconstruction noise alone. More metric details are provided in Appendix A.3.

Acquisition cost. Four GoPro cameras cost approximately $1500 USD. About 30% of sequences require alignment annotation, averaging 30 seconds per corrected sequence.

4 Downstream Evaluation

Our evaluation centers on whether the collected demonstrations support scene-aware large-object handling. The generation benchmark measures the value of explicitly conditioning on obstacles; the simulation study examines the use of interaction references on a robot embodiment; and obstacle replacement probes transfer beyond the modular capture layouts. These tests address different aspects of downstream utility and should not be interpreted as interchangeable evidence. In particular, reduced geometric penetration in generated motion is distinct from successful execution under simulated dynamics. The supplementary material additionally reports the unseen-object forecasting evaluation and implementation details.

4.1 Scene-constraint Locomotion Synthesis

This task tests whether a model can use SLoHa demonstrations to adapt large-object carrying motions to surrounding obstacles. The challenge is joint: keeping the body clear of an obstacle is insufficient if the carried object collides with it. The model must coordinate locomotion and object

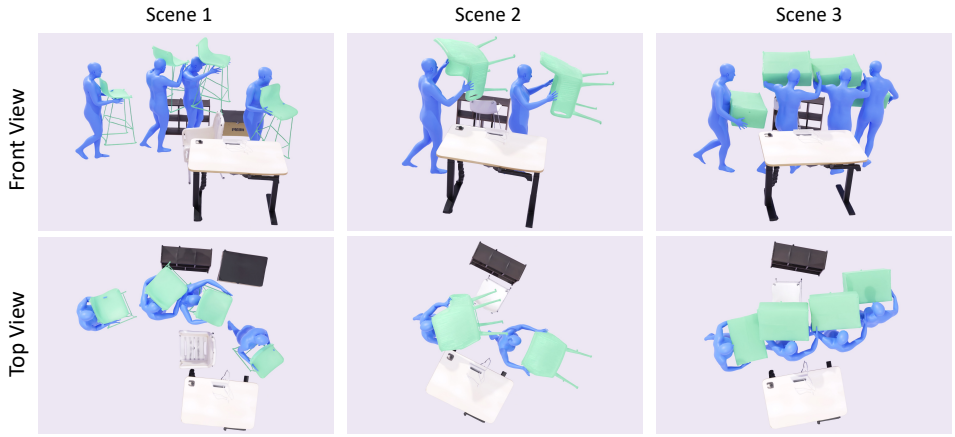


Figure 6: **Qualitative results of box-to-object obstacle replacement.**

Table 3: Quantitative results of locomotion in cluttered scenes.

	Human			Object		
Conditioning	$Pene_{\%scene} \downarrow$	$Pene_{mean} \downarrow$	$Pene_{max} \downarrow$	$Pene_{\%scene} \downarrow$	$Pene_{mean} \downarrow$	$Pene_{max} \downarrow$
w/o scene cond.	16%	1.26	7.55	27%	24.63	158.98
w/ scene cond.	**13%**	**0.78**	**4.69**	**18%**	**12.64**	**72.23**

orientation along a desired route, including local adjustments for constrained passages and turns. We generate human–object trajectories (H, O) from the initial human state h_0, object pose o_0, object mesh M, instruction T, sparse waypoints W, and scene occupancy S:

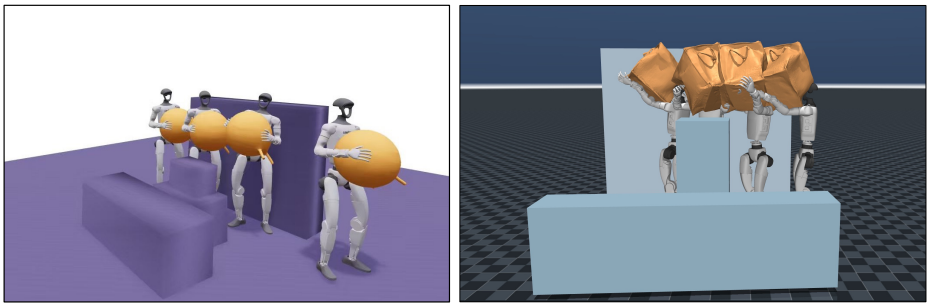
$$p_\theta(H, O \mid h_0, o_0, M, W, S, T).$$

Waypoints guide the route while leaving local body and object adjustments to the model. Sampling and implementation details are provided in Appendix A.6. Here, H is the sequence of human poses and O the corresponding sequence of object rotations and translations. The object mesh specifies the carried geometry, while scene occupancy specifies the surrounding barriers. Sparse route guidance does not prescribe a full pose sequence: the model must still determine how the body and object move between waypoints in response to the available clearance.

Baselines. Both settings use the same CHOIS (Li et al., 2024) backbone, SLoHa motions, and data split. **Without Scene Condition** omits scene geometry; **With Scene Condition** additionally receives reconstructed occupancy. This comparison isolates the contribution of scene input. CHOIS jointly generates human and object motion with a conditional diffusion model. We augment this backbone with occupancy-based scene conditioning while retaining the initial states, object geometry, instruction, and waypoint guidance in both settings. Thus, the no-scene model still receives task and route information, but cannot directly distinguish nearby obstacles from free space. The comparison concerns access to scene geometry, not a change in the source demonstrations. We split the data by sequence into 90% training and 10% testing, keeping adjacent clips from the same recording within one partition.

Metrics. Following DIMOS (Zhao et al., 2023), we separately measure human and object penetration: the proportion of vertices intersecting the scene ($\text{Pene}_{\%\text{scene}}$), mean penetration score ($\text{Pene}_{\text{mean}}$), and maximum score ($\text{Pene}_{\text{max}}$). These measures capture complementary aspects of scene compliance: the ratio reflects the extent of intersection, the mean summarizes penetration score, and the maximum exposes severe collisions that an average may conceal. Reporting the human and object separately is particularly important for large-object transportation, where their swept volumes differ substantially. Lower values are better; these geometric metrics do not by themselves establish dynamic stability.

Result. Scene conditioning reduces all reported human and object penetration metrics (Table 3). Qualitative comparisons in Fig. 11 show corresponding improvements in obstacle avoidance, supporting the value of scene input for large-object motion generation. For the carried object, mean penetration decreases from 24.63 to 12.64 and maximum penetration from 158.98 to 72.23. Human



(a) VideoMimic Retarget Result in Simulator (b) ResMimic Retarget Result in Simulator

Figure 7: Humanoid simulation results using SLoHa demonstrations. (a) VideoMimic retargeting and reference-motion tracking in simulation. (b) OmniRetarget–ResMimic retargeting and physics-based object carrying in simulation.

mean penetration also decreases from 1.26 to 0.78. The improvements on both components indicate that scene conditioning benefits the interaction trajectory rather than only the body's path. Residual penetration remains, so the results support improved scene compliance rather than a guarantee of collision-free execution. Generalization to unseen objects is evaluated separately in the supplementary material.

4.2 Humanoid Simulation

To assess SLoHa's utility for humanoid applications, we use two retargeting and control pipelines to demonstrate that its reconstructed motions can be transferred to a humanoid and executed in simulation. With VideoMimic (Allshire et al., 2025), we retarget SMPL-X motions to G1 and fine-tune the scene-conditioned tracking policy on SLoHa motions and scenes, improving aggregate tracking success from 4.88% to 93.20% (Appendix A.7.1, Table 6). In the second pipeline, we first use OmniRetarget (Yang et al., 2025) to retarget SMPL-X motions to G1, then use the retargeted motions as references to train a control policy with ResMimic (Zhao et al., 2025) in the simulator. Figure 7(a,b) shows representative results from the two pipelines. Training and evaluation details are provided in Appendix A.7. We further demonstrate real-world execution on a Unitree G1 humanoid, with qualitative results provided in Appendix A.8.

4.3 Box-to-Object Obstacle Replacement

The modular capture setup raises a practical question: do the learned carrying behaviors depend on box-shaped surroundings, or can they be used with other obstacle geometry? This experiment provides a qualitative check of the abstraction used during data acquisition. It changes the obstacle instances at test time while retaining the task of transporting a large object through constrained space. We test the SLoHa-trained CHOIS model from Sec. 4.1 in scenes with real objects replacing box obstacles, without additional training. Given the new scene and task, the model generates human and carried-object trajectories. Figure 6 provides qualitative evidence that the learned carrying behaviors can transfer beyond the modular box layouts. This result supports using simple physical barriers to collect useful motion adaptations, while remaining a limited qualitative test rather than evidence of generalization to arbitrary environments.

5 Limitation and Conclusion

We present SLoHa, a 4D HOI dataset of scene-constrained large-object handling, together with a democratic acquisition solution based on sparse multi-view videos and human-assisted reconstruction. Motivated by robotic transportation tasks, the dataset records the coupled human, object, and scene information needed to study obstacle-aware interactions. The scene-conditioned generation benchmark demonstrates improved geometric compliance. Since our current scenes are mainly corridor-style layouts, our reconstruction/obstacle representations may miss fine details under heavy occlusion, especially contact details. Future work will expand to more diverse environments and interactions, improve reconstruction fidelity and contact modeling, and explore closed-loop deployment on real humanoid robots.

AI USE STATEMENT

We used generative AI tools to assist with drafting and editing the manuscript, writing scripts for data processing, and aggregating and summarizing data. The authors take full responsibility for the final content of this work, including the text, code, data summaries, and claims produced with AI assistance.

ETHICS STATEMENT

All participants were adults and consented to participation, recording, research use, and the public release of the reconstructed data described below. This study did not obtain institutional ethics or IRB approval. Participants could stop at any time, and recording took place with another person present to supervise and assist. Before recording, participants tried lifting the objects; object weight and size were limited according to their feedback. The planned public dataset release is restricted to reconstructed SMPL-X motion, object data, and scene data, excluding raw RGB recordings. Any future release of RGB recordings, including the treatment of facial information, will be discussed separately with participants before release. Although excluding RGB recordings reduces direct visual identification, reconstructed body shape and motion may still carry identifying information. The dataset covers a limited set of participants and structured environments and should not be assumed to represent all people or real-world settings. Simulation-based evaluations do not establish the safety of deployment on physical robots.

REPRODUCIBILITY STATEMENT

Section 3 describes the capture setup and reconstruction pipeline. Appendix A.1 provides additional acquisition and data-processing details, and Appendix A.2 presents the interaction-alignment procedure and optimization objective. Section 4 describes the downstream tasks, baseline comparison, and evaluation metrics; Appendices A.6 and A.7 summarize the generation setup and humanoid retargeting and training procedure. We will publicly release the dataset and code upon acceptance to support replication and further research. The dataset release will comprise SMPL-X motion, object data, and scene data, rather than raw RGB recordings.

REFERENCES

Easymocap - make human motion capture easier. Github, 2021. URL https://github.com/zju3dv/EasyMocap.

Arthur Allshire, Hongsuk Choi, Junyi Zhang, David McAllister, Anthony Zhang, Chung Min Kim, Trevor Darrell, Pieter Abbeel, Jitendra Malik, and Angjoo Kanazawa. Visual imitation enables contextual humanoid control. In *Proceedings of The Conference on Robot Learning*, Proceedings of Machine Learning Research, 2025.

Joao Pedro Araújo, Jiaman Li, Karthik Vetrivel, Rishi Agarwal, Jiajun Wu, Deepak Gopinath, Alexander William Clegg, and Karen Liu. Circle: Capture in rich contextual environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21211–21221, 2023.

Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15935–15946, 2022.

Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, et al. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In *European Conference on Computer Vision*, pp. 557–577. Springer, 2022.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7291–7299, 2017.

Jiahao Chang, Chongjie Ye, Yushuang Wu, Yuantao Chen, Yidan Zhang, Zhongjin Luo, Chenghong Li, Yihao Zhi, and Xiaoguang Han. Reconviagen: Towards accurate multi-view 3d object reconstruction via generation. *arXiv preprint arXiv:2510.23306*, 2025.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, 2017.

Ke Fan, Shunlin Lu, Minyue Dai, Runyi Yu, Lixing Xiao, Zhiyang Dou, Junting Dong, Lizhuang Ma, and Jingbo Wang. Go to zero: Towards zero-shot motion generation with million-scale data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13336–13348, 2025.

Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10933–10942, 2021a.

Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *Int. J. Comput. Vision*, 2021b.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5152–5161, 2022.

Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2282–2292, 2019.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013.

Nan Jiang, Zimo He, Zi Wang, Hongjie Li, Yixin Chen, Siyuan Huang, and Yixin Zhu. Autonomous character-scene interaction synthesis from text instruction. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024a.

Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1737–1747, 2024b.

Jeonghwan Kim, Jisoo Kim, Jeonghyeon Na, and Hanbyul Joo. Parahome: Parameterizing everyday home activities towards 3d generative modeling of human-object interactions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1816–1828, 2025.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Chenghong Li, Hongjie Liao, Yihao Zhi, Xihe Yang, Zhengwentai Sun, Jiahao Chang, Shuguang Cui, and Xiaoguang Han. Mvhumannet++: A large-scale dataset of multi-view daily dressing human captures with richer annotations for 3d human digitization. *arXiv preprint arXiv:2505.01838*, 2025.

Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023.

Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object interaction synthesis. In *European Conference on Computer Vision*, pp. 54–72. Springer, 2024.

Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36:25268–25280, 2023.

Yun Liu, Chengwen Zhang, Ruofan Xing, Bingda Tang, Bowen Yang, and Li Yi. Core4d: A 4d human-object-human interaction dataset for collaborative object rearrangement. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1769–1782, 2025.

Jiaxin Lu, Chun-Hao Paul Huang, Uttaran Bhattacharya, Qixing Huang, and Yi Zhou. Humoto: A 4d dataset of mocap human object interactions. *arXiv preprint arXiv:2504.10414*, 2025.

Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5442–5451, 2019.

CMU Mocap. Cmu graphics lab motion capture database., 2003.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10975–10985, 2019.

Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 722–731, 2021.

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.

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.

Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European conference on computer vision*, pp. 581–600. Springer, 2020.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025.

Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. *Advances in Neural Information Processing Systems*, 35:14959–14971, 2022.

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17868–17879, 2024.

Zhangyang Xiong, Chenghong Li, Kenkun Liu, Hongjie Liao, Jianqiao Hu, Junyi Zhu, Shuliang Ning, Lingteng Qiu, Chongjie Wang, Shijie Wang, et al. Mvhumannet: A large-scale dataset of multi-view daily dressing human captures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19801–19811, 2024.

Liang Xu, Chengqun Yang, Zili Lin, Fei Xu, Yifan Liu, Congsheng Xu, Yiyi Zhang, Jie Qin, Xingdong Sheng, Yunhui Liu, et al. Perceiving and acting in first-person: A dataset and benchmark for egocentric human-object-human interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12535–12548, 2025.

Lujie Yang, Xiaoyu Huang, Zhen Wu, Angjoo Kanazawa, Pieter Abbeel, Carmelo Sferrazza, C. Karen Liu, Rocky Duan, and Guanya Shi. Omniretarget: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction. *arXiv preprint arXiv:2509.26633*, 2025.

Hongwei Yi, Justus Thies, Michael J Black, Xue Bin Peng, and Davis Rempe. Generating human interaction motions in scenes with text control. In *European Conference on Computer Vision*, pp. 246–263. Springer, 2024.

Juze Zhang, Haimin Luo, Hongdi Yang, Xinru Xu, Qianyang Wu, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Neuraldome: A neural modeling pipeline on multi-view human-object interactions. In *CVPR*, 2023.

Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Hoi-mˆ 3: Capture multiple humans and objects interaction within contextual environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 516–526, 2024.

Chengfeng Zhao, Juze Zhang, Jiashen Du, Ziwei Shan, Junye Wang, Jingyi Yu, Jingya Wang, and Lan Xu. I'm hoi: Inertia-aware monocular capture of 3d human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 729–741, 2024.

Kaifeng Zhao, Yan Zhang, Shaofei Wang, Thabo Beeler, and Siyu Tang. Synthesizing diverse human motions in 3d indoor scenes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 14738–14749, 2023.

Siheng Zhao, Yanjie Ze, Yue Wang, C. Karen Liu, Pieter Abbeel, Guanya Shi, and Rocky Duan. Resmimic: From general motion tracking to humanoid whole-body loco-manipulation via residual learning. *arXiv preprint arXiv:2510.05070*, 2025.

A Supplementary Material

A.1 Additional Dataset and Capture Details

Capture system. Our multi-view capture system consists of five GoPro cameras: four static exocentric cameras placed around the scene and one egocentric camera mounted on the subject's head. The cameras jointly cover an interaction area of approximately 10m^2 and record 4K videos at 30 FPS. The current SLoHa reconstruction pipeline uses only the four exocentric views; the egocentric recordings are retained but are not used in this paper.

Calibration and synchronization. Whenever a scene layout is changed or a camera is moved and re-mounted, we place a calibration board in the scene to estimate the multi-view geometry and remove it before recording the interaction. All cameras are synchronized using timecode to obtain frame-level temporal alignment. VGGT (Wang et al., 2025) initially reconstructs the scene in its own coordinate system. We use the corresponding camera extrinsics from calibration and VGGT to estimate a transformation matrix from the reconstruction coordinate system to the calibrated real-world coordinate system. After aligning the object mesh scale with the VGGT point cloud as described below, we express the reconstructed human motion, object geometry and trajectories, and scene geometry in this common world frame.

Collection protocol. Each task instruction specifies the target object, transportation goal, and safety constraints. A representative instruction is: "Hold the umbrella storage bucket with both hands, walk through the passageway, and do not hit the obstacles." Subjects are asked to move naturally without exaggerating obstacle-avoidance behavior. The instruction fixes the task but leaves the choice of posture, foot placement, and object orientation to the subject. This distinction is important for collecting scene-induced adaptations: a narrow passage or an overhead barrier changes how the task is performed, rather than simply providing a background for an otherwise unchanged motion. Cardboard boxes make these constraints easy to rearrange across recording sessions. For each room configuration, two subjects receive approximately one day of familiarization. Collection spans roughly three months using two capture systems deployed in two rooms with consistent camera placement.

Scene reconstruction. We use VGGT (Wang et al., 2025) to estimate per-view depth maps, point maps, and camera parameters from synchronized RGB views. Because the task is governed primarily by obstacle occupancy, each structured obstacle is represented with a 3D bounding box encoding its height, width, and available clearance. We first obtain 2D obstacle proposals using Grounded-SAM-2 (Ren et al., 2024). Humans verify the proposals; for failures, an annotator supplies point prompts and SAM2 generates corrected masks. Verified masks are lifted into 3D using the estimated camera parameters and depth maps, from which tight obstacle bounding boxes are computed.

Object scale and pose. We reconstruct a reusable object mesh with ReconViaGen (Chang et al., 2025) and track its masks with SAM2 (Ravi et al., 2024). Using the object masks, we extract the object points from the VGGT point cloud. We compute bounding spheres for this object point cloud and the reconstructed mesh, with diameters d_{pc} and d_{mesh}, respectively. The ratio $s = d_{\text{pc}}/d_{\text{mesh}}$ provides a uniform scale factor for resizing the mesh to match the VGGT reconstruction. This step resolves the relative scale between object-level and scene-level reconstruction; the calibration-based coordinate transformation then places the reconstructed data in the real-world frame. FoundationPose (Wen et al., 2024) estimates the per-frame 6D pose from the rescaled mesh, object masks, depth maps, and camera parameters expressed consistently in the same frame. Mask propagation helps maintain object association during temporary occlusions.

Object and layout design. The seven object categories are Sports/Toys, Storage, Tables, Seating/Bedding, Household Tools, Racks/Carts, and Decor. Object geometry affects grasp placement, object orientation, torso stabilization, and foot placement. The object set therefore spans elongated, wide, tall, and irregular shapes rather than relying on category diversity alone. Scene layouts elicit complementary adaptations: narrow passages encourage body alignment; corners require turning and object re-orientation; overhead barriers prompt crouching; ground-level obstacles require high stepping; and lift-and-set-down scenes capture the approach, lift, carry, and placement sequence.

A.2 Interaction Alignment Details

We correct object-pose estimates using sparse multi-view annotations followed by contact-aware refinement.

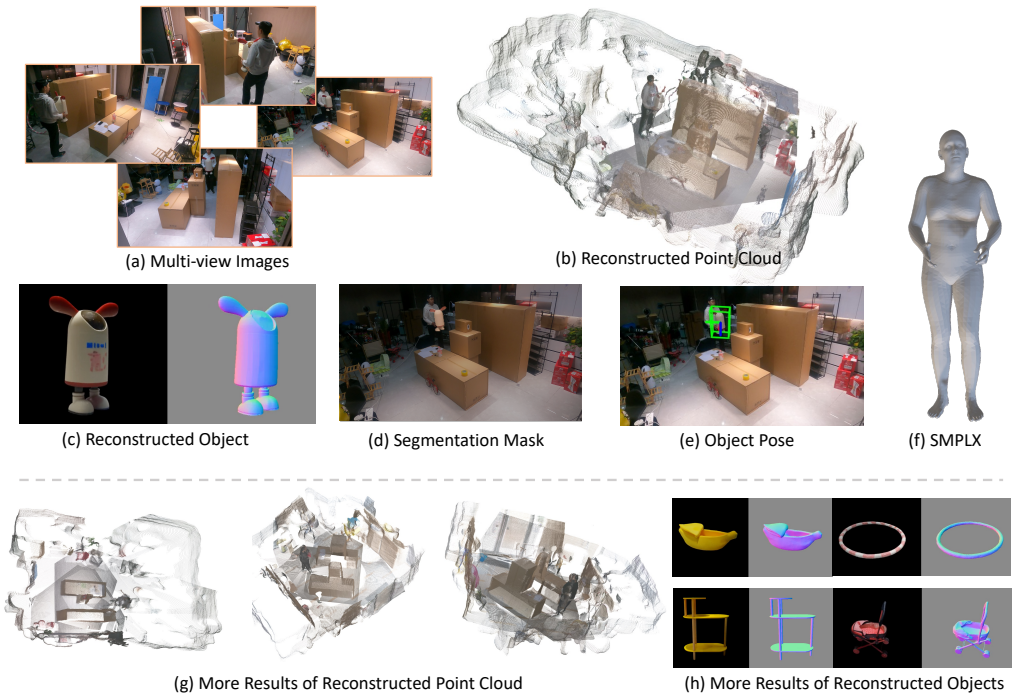


Figure 8: Additional intermediate results from the SLoHa data-processing pipeline.

Stage 1: 2D-guided pose initialization. We render the reconstructed interaction in each camera view and compare it with the recorded videos. When misalignment is detected, annotators mark 5–10 salient object points and their corresponding locations in the multi-view RGB frames. Calibrated cameras and VGGT depth/point maps lift these correspondences to 3D. We align the reconstructed object mesh to the scene-derived object point cloud to refine the first-frame pose, which initializes subsequent frames.

Stage 2: contact-aware pose refinement. In the reference frame, we associate each of K object surface points $\mathbf{o}_k$ with its nearest hand surface point $\mathbf{h}_k$. For each target frame, we optimize the object pose $(\mathbf{R}, \mathbf{t})$ using the original objective:

$$\min_{\mathbf{R},\mathbf{t}} \frac{1}{K} \sum_{k=1}^{K} \|\mathbf{R}\mathbf{o}_k + \mathbf{t} - \mathbf{h}_k\|_2^2 + \lambda_t \|\mathbf{t} - \mathbf{t}_{\text{net}}\|_2^2 + \lambda_R \|\mathbf{R} - \mathbf{R}_{\text{net}}\|_2^2, \tag{1}$$

Here, $\mathbf{R}$ and $\mathbf{t}$ map object coordinates to the world frame, and $(\mathbf{R}_{\text{net}}, \mathbf{t}_{\text{net}})$ are network-predicted poses. The weights λ_t and λ_R balance contact alignment against regularization toward the prediction.

A.3 Motion-quality Protocol

Table 4 extends the main-text motion comparison with interaction metrics and ParaHome. We follow the metric definitions in HUMOTO Appendix A.2 (Lu et al., 2025). Human- and object-motion metrics. Following Lu et al. (2025), *foot sliding* measures horizontal foot displacement near the ground, with lower values indicating less sliding during ground contact. *Jerk*, evaluated for both human and object motion, measures changes in acceleration; lower values indicate smoother trajectories. *MSNR* compares the energy of joint velocities with their residual relative to a temporally smoothed signal. Higher values indicate smoother motion, but may also reflect excessive smoothing rather than more accurate reconstruction.

Coherence measures the compactness of pose clusters: higher scores indicate poses that are more tightly grouped around their cluster centers. *Diversity* is the normalized entropy of the distribution

Table 4: Motion and interaction quality. Baseline values are reported by HUMOTO (Lu et al., 2025) (arXiv v2, Table 1). Bold and underlining indicate the best and second-best values, respectively, in each directional column among the datasets shown; MSNR has no monotonic preference. Contact entropy and state consistency describe interaction-state statistics, not contact accuracy alone.

Dataset	Human motion					Object motion	Contact		
	Foot sliding (cm) ↓	Jerk (m/s^3) ↓	MSNR (dB)	Coherence ↑	Diversity ↑	Jerk (m/s^3) ↓	Penetration (cm) ↓	Contact entropy ↑	State consistency ↑
OMOMO	2.130	15.10	12.37	<u>0.619</u>	<u>0.978</u>	27.40	<u>0.0602</u>	**1.9468**	0.4837
ParaHome	3.008	9.19	1.82	0.592	**0.980**	**0.08**	0.2167	1.0254	<u>0.6815</u>
HUMOTO	<u>0.958</u>	**1.87**	9.42	**0.653**	0.956	1.13	**0.0068**	<u>1.4587</u>	0.5061
SLoHa (Ours)	**0.658**	<u>4.33</u>	21.92	0.522	0.971	<u>0.76</u>	0.2003	1.4352	**0.839**

over pose clusters, with higher values indicating broader coverage of motion patterns. Both depend on the motion content and clustering representation; neither alone distinguishes meaningful motion variation from reconstruction noise.

Interaction metrics. *Penetration* measures the depth of object intrusion into the human mesh, with lower values indicating less geometric intersection. *Contact entropy* measures the diversity of transitions between interaction states, including penetration, contact, proximity, and separation. Higher entropy indicates more varied state transitions, not necessarily more accurate contacts. *State consistency* measures how persistently these states are maintained over time, with a penalty for prolonged large penetration. Higher consistency indicates fewer state changes, but does not guarantee correct contact geometry. The metrics should therefore be interpreted jointly and in the context of each dataset's tasks.

Capture and processing context. HUMOTO is designed for high-fidelity interaction capture, combining specialized motion-capture suits and gloves with RGB-D observations, precisely artist-modeled objects, and professional manual cleaning and verification (Lu et al., 2025). These hardware and processing requirements entail greater resource demands than our lightweight capture and sparse-alignment approach. Its lowest penetration in Table 4 is consistent with this emphasis on detailed contact fidelity and explicit cleanup of intersections. SLoHa instead prioritizes accessible acquisition and coverage of scene-constrained large-object motion; this difference in design goals contextualizes, but does not eliminate, its contact-geometry limitations.

Motion quality. SLoHa has the lowest foot sliding among the four datasets and lower object jerk than OMOMO and HUMOTO, though not ParaHome, while retaining a diversity score close to OMOMO and ParaHome. Its human-motion jerk is lower than OMOMO and ParaHome but higher than HUMOTO, and its coherence is lower than all three. These differences may partly reflect the varied postural adjustments required for scene-constrained carrying, but this comparison alone cannot separate task complexity from reconstruction error. Similarly, low jerk or high MSNR does not by itself establish accurate motion recovery.

Interaction quality and scope. SLoHa exhibits higher penetration than OMOMO and HUMOTO, although slightly lower than ParaHome, highlighting remaining limitations in contact geometry. Its contact entropy is lower than OMOMO and HUMOTO but higher than ParaHome, while state consistency is the highest among the compared datasets. Relatively persistent interaction states are consistent with sustained object carrying, which may involve fewer state transitions than repeated grasping and releasing. SLoHa is therefore intended to support scene-constrained whole-body and object-motion modeling, with limitations in fine-grained contact fidelity. The downstream generation and simulation experiments assess its utility through the corresponding learning pipelines, rather than establishing the accuracy of the original contacts.

A.4 Qualitative Comparison with a Mocap System

To qualitatively assess the captured data, we compare our reconstruction with a marker-based Nokov motion capture system. Objects are placed inside the mocap area with markers attached, and subjects wear mocap suits. Four GoPro cameras record the same sequences using our capture setup.

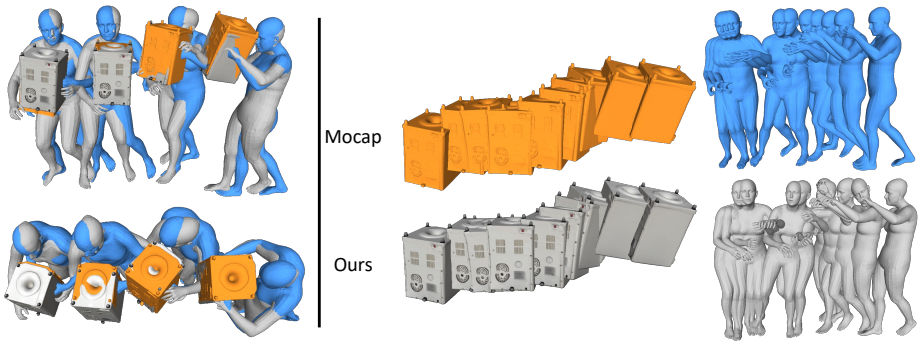


Figure 9: We visualize our capture results (uncolored) and mocap results (colored) in the aligned coordinate system. The object trajectories are well aligned, while human poses and hand-object contacts show moderate differences due to the limitations of both capture setups.

Table 5: Quantitative results of unseen large-object handling.

Training Dataset	$C_prec \uparrow$	$C_recall \uparrow$	$C_F1 \uparrow$	$C\% \uparrow$	$FID \downarrow$	$FS \downarrow$
OMOMO (Li et al., 2023)	0.31	0.27	0.30	0.34	1.60	0.38
Ours	**0.38**	**0.32**	**0.34**	**0.37**	**1.25**	**0.13**

We collect and process 10 sequences and align the coordinate systems of the two systems for comparison. As shown in Fig. 9, the reconstructed object trajectories are visually consistent with the mocap results. Some differences remain in human poses and hand-object contacts. The mocap setup is primarily designed for body-pose capture, with cameras relatively far from the hands, making fine hand motion difficult to capture. In our reconstruction, the black mocap suits may also affect SMPL-X fitting and contact quality. This comparison provides qualitative evidence rather than a quantitative reconstruction-accuracy measurement.

A.5 Application to Unseen Objects

This complementary evaluation tests application to unseen carried-object geometry. The goal is to predict future human and object trajectories from the initial human/object states, object geometry, and sparse waypoint guidance. We test the trained models on unseen objects (Fu et al., 2021b), assessing whether the learned large-object interaction dynamics transfer to novel object geometries. Specifically, given the initial human state h_0, the initial object pose $o_0 \in SE(3)$, the object mesh M, the text instruction(e.g., "move the large box") T, and sparse waypoints W that indicate the planned route, the model predicts future human and object trajectories. The task is defined as:

$$p_\theta(H, O \mid h_0, o_0, M, W, T),$$

where H and O denote the forecasted human and object motion sequences over the prediction horizon. We use the same CHOIS (Li et al., 2024) backbone to train on our dataset and the OMOMO dataset (Li et al., 2023), respectively. Then test both models on the same unseen objects to compare their ability to handle novel objects.

Metrics. We evaluate both interaction correctness and motion realism using a set of standard metrics following Li et al. (2024). Contact accuracy is measured by contact precision ($C_$prec), contact recall ($C_$recall), and contact F1-score (C_F1). In addition, we report the contact percentage ($C\%$), defined as the proportion of frames in which human–object contact is detected, to quantify how consistently the predicted motions maintain interaction. To assess overall motion quality, we compute FID, which measures the discrepancy between the distributions of generated motions and ground-truth motions in a learned feature space. Finally, we report the foot-sliding score (FS) in centimeters, following Li et al. (2024), which measures the weighted average of foot displacement in the horizontal plane; lower values indicate less sliding.

Table 5 shows that the model trained on SLoHa achieves higher contact precision, recall, and F1-

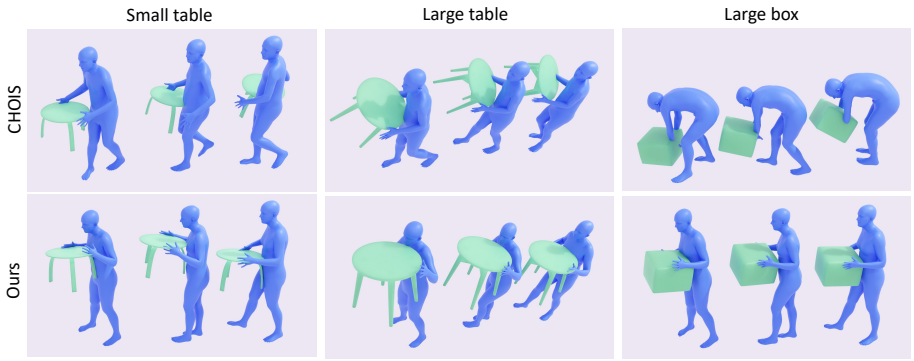


Figure 10: **Unseen results visualization.**Visualization of human–object motion forecasting on novel large objects using the same CHOIS backbone trained on different datasets. The text instruction for this test is "move the {object name}".

score and a lower FID than the model trained on OMOMO. Qualitative results in Fig. 10 provide additional examples of generated interactions with novel objects. These results support the utility of SLoHa for unseen-object interaction modeling.

A.6 Scene-aware Generation Details

The model receives the initial SMPL-X state, object pose and mesh, a task instruction, and sparse 2D waypoints $W = \{(x_k, y_k)\}_{k=1}^{K}$ sampled every n frames along the planned route. These waypoints constrain global navigation while leaving local turning, bending, and object reorientation to the model. Both ablations use the same CHOIS backbone and data split; only the scene-conditioned setting receives reconstructed occupancy. The CHOIS processing protocol treats hands as rigid parts rather than modeling articulated fingers.

Scene conditioning and data split. We follow the occupancy-based scene-encoding approach of LINGO (Jiang et al., 2024a), using a scene encoder to extract features from the reconstructed occupancy representation and supply them as an additional conditioning signal to the CHOIS diffusion model. We split the data at the sequence level, using 90% for training and 10% for testing, with no adjacent clips from the same recording shared across splits. The scene-conditioned and no-scene baselines use the same partition; scene conditioning is omitted in the latter.

A.7 Humanoid Simulation: Retargeting and Training

A.7.1 VideoMimic Retarget Experiment

Training. We adopt VideoMimic (Allshire et al., 2025) to test whether SLoHa motions and reconstructed scenes support humanoid retargeting in cluttered environments in a simulator. We retarget the SMPL-X motion sequences to the G1 humanoid and initialize the policy from the released VideoMimic mocap pre-training checkpoint. Only the scene-conditioned tracking policy is further trained. During this fine-tuning stage, AMASS data are disabled and training uses SLoHa motion clips with their corresponding reconstructed scene geometry. The resulting policy learns to follow the retargeted motion references while moving through the simulated scenes.

Evaluation. We evaluate in 24 scenes with box obstacles and 10 scenes with real objects as obstacles, with 676 episodes for each training setting. An episode is successful if the robot completes the rollout while remaining within a predefined distance threshold of the reference trajectory throughout the episode. Failures are categorized as falling (upper-body contact with the ground), obstacle blockage (physical obstruction causing the reference-distance threshold to be exceeded), or unstable motion (loss of stable locomotion exceeding the threshold without direct obstacle blockage).

Results. Table 6 compares the AMASS-trained VideoMimic baseline without SLoHa fine-tuning against the SLoHa-trained policy. Aggregate success increases from 33/676 (4.88%) to 630/676 (93.20%), while obstacle-blockage failures decrease from 604 to 39. These results support the

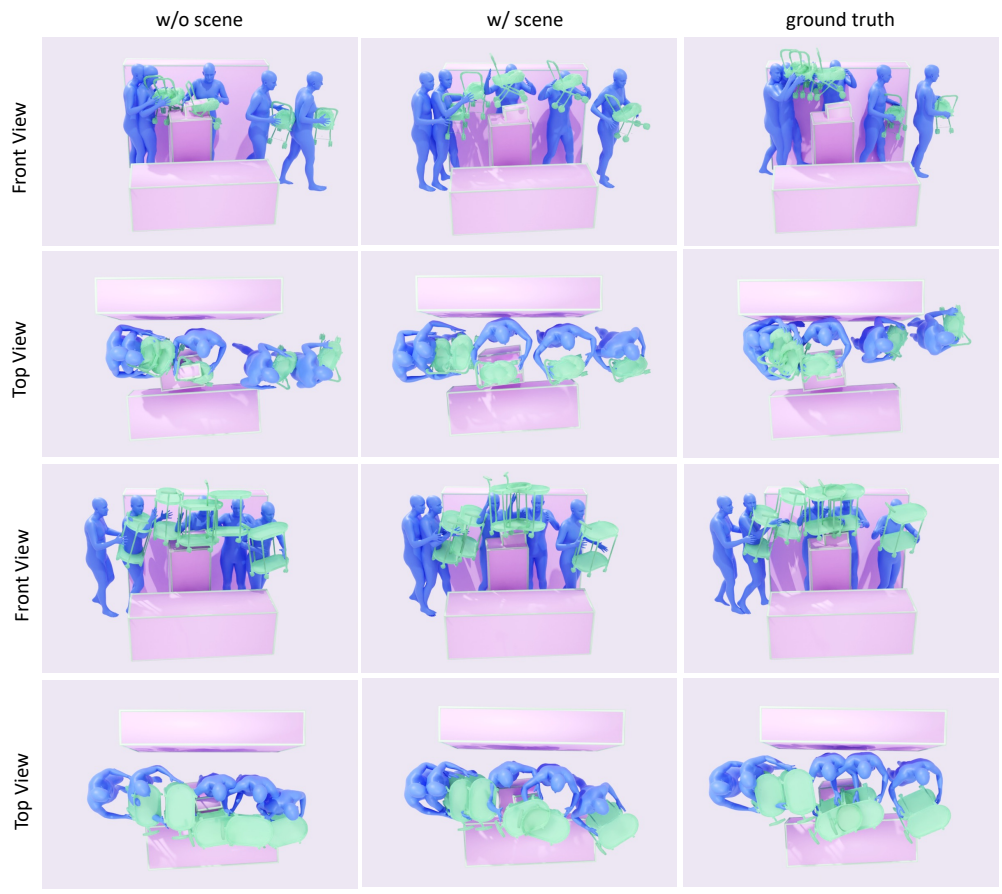


Figure 11: **Qualitative comparison and visualization of results with and without scene conditioning.**

Table 6: VideoMimic reference-tracking results, aggregated over box- and real-object-obstacle scenes.

			Failure		
Training data	Episodes	Success	Falling	Obstacle blockage	Unstable motion
AMASS	676	33/676 (4.88%)	10 (1.48%)	604 (89.35%)	29 (4.29%)
SLoHa	676	630/676 (93.20%)	5 (0.74%)	39 (5.77%)	2 (0.30%)

utility of SLoHa motion and scene references for this tracking task. Qualitative examples are shown in Fig. 7(a). The table aggregates both obstacle settings and does not provide separate success rates for each setting.

A.7.2 ResMimic Retarget Experiment

Retargeting. We use time-aligned SMPL-X motion and object trajectories in a common world coordinate system. Following OmniRetarget (Yang et al., 2025), we optimize robot poses to preserve local body–object interaction geometry, subject to joint-limit, foot-contact, and non-penetration constraints. The object trajectory remains fixed during retargeting, producing synchronized robot and object references.

Policy training. In Isaac Gym, we adopt Resmimic (Zhao et al., 2025). A frozen general motion-tracking policy supplies nominal actions, and a task-specific residual policy adds object-aware corrections. The residual actor receives robot observations, robot-motion references, and the current

object pose; the reference object trajectory enters the reward and termination criteria rather than the actor input. We train a separate residual policy for each sequence using PPO, with 512 parallel environments for up to 5,000 iterations. Rewards encourage robot-motion tracking, object-trajectory tracking, and hand support, with penalties for undesirable contacts and unstable motion. The base policy remains frozen, and checkpoints are selected by inspecting complete rollouts for motion tracking, contact, and balance.

Physical interaction. Visual and collision geometries share a consistent object frame and scale. During policy rollouts, the object is a dynamic rigid body subject to gravity and contact, rather than replayed along its reference trajectory. We use neither kinematic attachment nor grasp assistance or object-following forces, so object state depends on physical robot–object contact.

A.8 Real-world Experiments on Unitree G1

To complement the simulation experiments, we evaluate the ResMimic-based pipeline on a physical Unitree G1 humanoid. As described in Appendix A.7.2, SLoHa SMPL-X demonstrations are first retargeted to G1 using OmniRetarget, and the resulting references are used for policy training with ResMimic in simulation. The real-world experiments provide a qualitative check of whether the learned carrying behaviors can also be executed on physical hardware.

Figure 12 presents two recorded real-world executions, with six frames from each video ordered chronologically. These examples complement the simulated rollouts with hardware evidence for transferring SLoHa-derived behaviors through the retargeting and control pipeline.

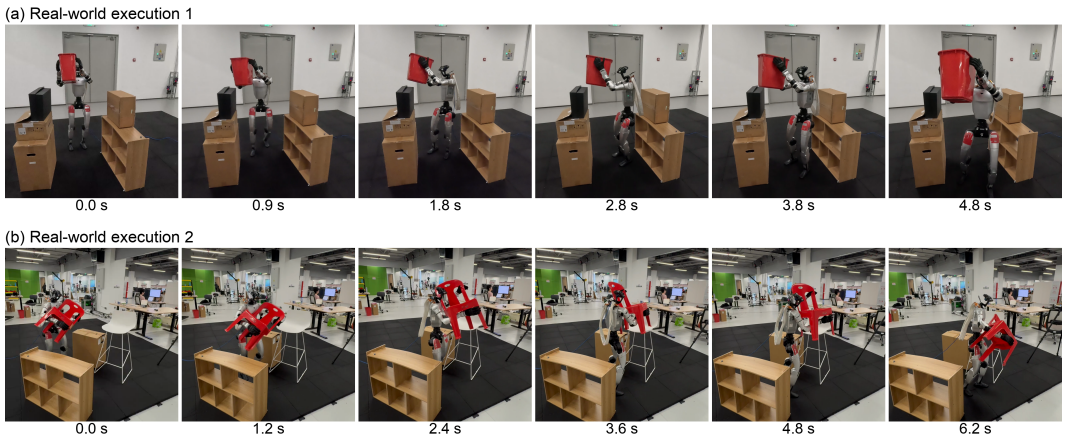


Figure 12: Real-world execution using ResMimic on a Unitree G1 humanoid. Rows (a) and (b) show two recorded experiments; each row contains six chronological frames, with timestamps measured from the start of its source video.