HUMOTO: A 4D Dataset of Mocap Human Object Interactions

Jiaxin Lu[1,2*], Chun-Hao Paul Huang[2], Uttaran Bhattacharya[2], Qixing Huang[1], Yi Zhou[2]
[1]University of Texas at Austin, [2] Adobe Research
lujiaxin@utexas.edu, {chunhaoh, ubhattac, yizho}@adobe.com, huangqx@cs.utexas.edu

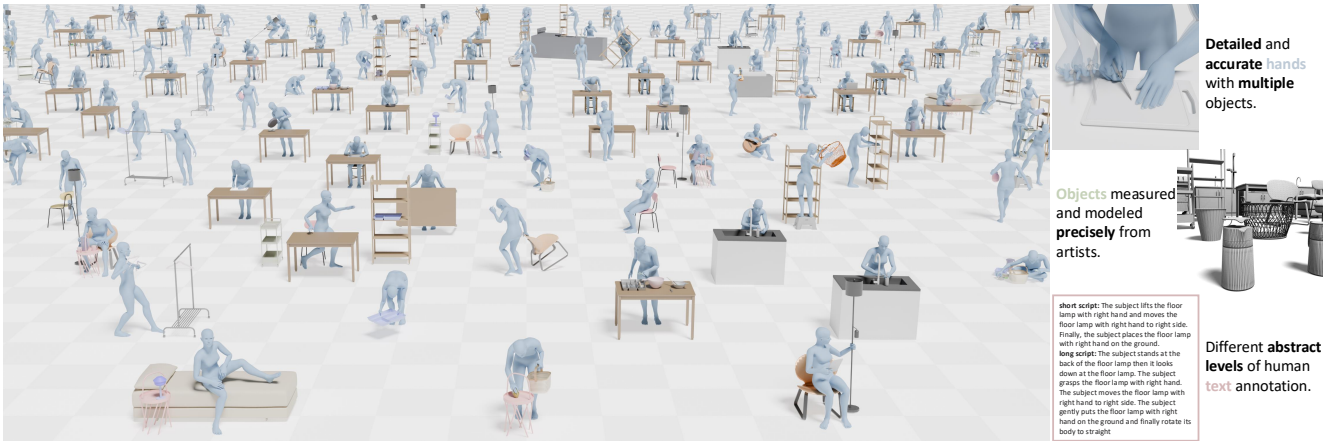


Figure 1. **Overview of the HUMOTO dataset.** The dataset contains mocap 4D human-object interaction animations with multiple objects. The unique features of the dataset include its detailed, accurate interaction modeling, specifically the detailed hand pose. The objects are precisely modeled by artists. We additionally provide different abstract levels of text annotation for the interactions.

Abstract

We present Human Motions with Objects (HUMOTO), a high-fidelity dataset of human-object interactions for motion generation, computer vision, and robotics applications. Featuring 736 sequences (7,875 seconds at 30 fps), HUMOTO captures interactions with 63 precisely modeled objects and 72 articulated parts. Our innovations include a scene-driven LLM scripting pipeline creating complete, purposeful tasks with natural progression, and a mocap-and-camera recording setup to effectively handle occlusions. Spanning diverse activities from cooking to outdoor picnics, HUMOTO preserves both physical accuracy and logical task flow. Professional artists rigorously clean and verify each sequence, minimizing foot sliding and object penetrations. We also provide benchmarks compared to other datasets. HUMOTO's comprehensive full-body motion and simultaneous multi-object interactions address key data-capturing challenges and provide opportunities to advance realistic human-object interaction modeling across research domains with practical applications in animation, robotics, and embodied AI systems. Project: https://jiaxin-lu.github.io/humoto/.

1. Introduction

4D Human-Object Interaction (HOI) data are crucial for understanding human behaviors in our three-dimensional world and for numerous applications in computer vision [41, 49, 50, 69, 75, 76, 78], robotics [4, 12, 46, 48, 52, 62], computer graphics [25, 32, 56], and generative AI [2, 26, 40, 71]. These applications range from HOI detection and reconstruction to motion generation, robotic learning through demonstration, and even image/video generation. All of these fields rely on 4D HOI data to capture human and object poses, ground-truth geometries, dynamics, forces, and multi-view observations [70, 72]. However, the lack of realistic 4D data hampers progress, particularly in scenarios involving multiple objects and detailed manipulations [16, 35]. As both generative and discriminative models advance [10, 40, 49, 51, 67], the need for high-quality HOI data has become increasingly critical.

Acquiring high-quality 4D HOI data is expensive due to

* The work was mainly conducted at Adobe Research.

the need for sophisticated motion capture setups and extensive manual data cleaning. Although recent efforts have provided various 4D human-object motion datasets [1, 35, 39, 58, 59, 61, 77, 80, 81], most focus on single-object interactions or lack detailed hand movements. Comprehensive datasets that capture interactions with multiple objects, with full-body and hand motion, remain a gap in the field.

To address this, we introduce Human Motions with Objects (HUMOTO), a new 4D animation dataset captured from real performance. HUMOTO includes 736 curated sequences totaling 7,875 seconds of motion (captured at 30 fps), featuring diverse daily activities and interactions with 63 objects comprising 72 distinct parts. Many scenes involve interactions with multiple objects, such as meal preparation with various utensils, storage organization, and room arrangement. The objects span a wide range of sizes, from small household items like utensils and tools to larger furniture pieces, all modeled based on real-world measurements. All human motions are captured with detailed body and hand movements, accompanied by text annotations.

The acquisition of HUMOTO is particularly challenging due to the complexity of recording fine-grained, multi-object interactions. It requires precise calibration, specialized equipment, and extensive post-processing to produce clean, high-quality sequences. By leveraging state-of-the-art techniques, including Large Language Model (LLM)-generated scripts and multi-sensor tracking, we create a dataset with unprecedented detail and fidelity.

Our dataset's distinctive quality stems from our complementary capture approach. To generate diverse motion scripts covering varied daily activities, we use a directorial mindset to design stories and actions, and we introduce a *Scene-Driven LLM Scripting* method to hierarchically generate these scripts. To capture human motion in the presence of frequent object occlusions, we utilize motion capture suits and gloves with electromagnetic field (EMF) technology to track performers, while dual-Kinect RGB-D sensors record object poses. This multi-modal system ensures fidelity in both large-scale movements and fine manipulations, even in occlusion-heavy scenarios.

All sequences undergo rigorous cleaning and independent verification by professional artists, with particular attention to common issues such as foot sliding and object penetration, ensuring clean yet natural movement nuances preserved data for machine learning context. An independent group of artists were also invited to assess the complete dataset's quality from a professional perspective. Moreover, we introduce a set of metrics to evaluate our and other HOI datasets, providing a comprehensive benchmark for human motion, object motion, and interaction quality.

HUMOTO provides a valuable resource for training models in motion generation, robotics, computer vision, and 2D generation. These sequences capture not only physical dynamics but also the logical progression of tasks, making them useful for learning natural action sequences and task planning [39, 48]. The comprehensiveness of the data set extends its utility in multiple domains: motion generation models can learn natural interaction patterns [10, 51], robotics researchers can study human manipulation strategies [13, 48, 52], and computer vision systems can train on accurate 3D ground truth for detection, tracking, and reconstruction [41, 50, 69, 78]. Image or video generation systems can also use verified motion sequences for content creation and authorization [25, 36, 71].

The contributions of this work include the following.

- A high-fidelity HOI dataset featuring complex, meaningful, and diverse daily interactions with multiple objects at various scales.
- A multi-modal capture methodology with Scene-Driven LLM Scripting and multi-sensors setup, preserving subtle interactions even in challenging occlusion scenarios.
- A set of quality metrics and benchmarks to evaluate HOI datasets to establish quantitative standards for human motion, object motion, and interaction quality.

2. Related Work

Human Motion Capturing Technologies. Recent advances in **human pose estimation from cameras**, including monocular RGB and RGB-D setups, have significantly broadened the scope of human motion capture. Early research explored markerless systems [5, 7, 14, 15, 34, 57], while more recent frameworks such as OpenPose [3] and DensePose [20] provide robust 2D and 3D joint detection. These camera-based systems are frequently improved using optimization techniques [23] or pre-trained models [54, 81], which substantially improve tracking accuracy. In parallel, marker-based pose estimation methods have been successfully applied to human-object interaction scenarios [44, 45, 47], delivering superior precision in specific contexts. Although these techniques are effective in unconstrained environments, they often encounter limitations when dealing with complex poses or occlusions.

Motion capture suits (mocap) have emerged as a widely adopted tool for capturing high-fidelity human motion across both research and industry applications. Both optical mocap systems and electromagnetic field suits have been employed in dataset collection [9, 17, 24, 37], offering extensive coverage for more challenging scenarios.

For **object pose estimation**, RGB-D cameras have become increasingly prevalent in HOI scenarios. Advanced techniques [42, 60, 66, 68] have demonstrated remarkable performance in object detection and localization. In the domain of neural systems, inertial measurement units (IMUs) have been attached to objects to track specific parameters [22, 79, 81].

Human-Object Motion Datasets. The field of human-

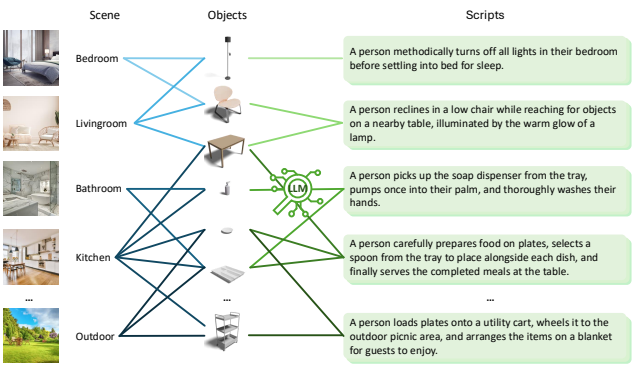


Figure 2. **Scene-Driven LLM Scripting.** We established target scenes, prepared relevant interaction objects, and then leveraged LLMs to generate detailed action scripts.

object interaction has witnessed the development of several significant datasets, each addressing different aspects of HOI capture. GRAB [58] represents one of the first data sets that addresses the full-body human-object interaction; however, it focuses primarily on upper-body interactions and is therefore omitted from our comparison. BEHAVE [1] and OMOMO [39] present more complex scenarios but lack detailed hand pose information. IMHD [77] specifically targets highly dynamic human-object interactions such as sports activities. Home [35] and TRUMANs [33] investigate human-object interactions within domestic environments, though these scenes tend to be more stationary with limited variance. TACO [44] focused more on capturing ego-centric interactions. Beyond dedicated data sets on human-object interaction, MIXAMO [29] provides a comprehensive repository of motion capture data used primarily in character animation and game development. HUMAN3.6M [82] constitutes a large-scale dataset designed for human motion capture, focusing on natural daily activities rather than human-object interactions.

While each of these datasets has significantly advanced the field, all exhibit limitations in capturing the complexity of real-world multi-object interactions. A critical shortcoming is the frequent inaccuracy of hand-object interactions, where hands either appear completely detached from objects or penetrate surfaces by significant margins. Additionally, many existing datasets consist of isolated, purposeless movements that, even with textual annotations, make it difficult to extract meaningful representations of continuous human daily activities. These limitations impede the development of models capable of understanding natural human-object interactions, particularly when involving multiple objects or requiring fine-grained manipulations.

3. Data Collection

The HUMOTO dataset advances human-object interaction research through a comprehensive collection methodology

Figure 3. **Capture environment.** *Left:* Overview of our capturing environment showing two Kinect cameras, stage, lighting, calibration board, and interaction objects. *Right:* Calibration procedure with the performer in a standardized position, enabling precise alignment between mocap suit data and camera coordinates.

Figure 4. **3D Meshes.** Artist-modeled objects used in HUMOTO.

that mirrors cinematic production. Beginning with LLM-generated scripts describing natural daily activities, we carefully selected and modeled common household objects before capturing interactions on a custom motion capture stage equipped with dual Kinect cameras.

3.1. Script Development

To address the limitations of existing datasets, where interactions often appear arbitrary or disconnected, we develop a systematic approach to create action scripts before capturing a large volume of motion data. Inspired by movie production workflows of grouping actions into scenes, we established a *Scene-Driven LLM Scripting* framework to automate script generation. First, we created conceptual "rooms" by logically grouping related objects from our collection. We then provided these groupings to LLMs to generate cohesive interaction sequences within contextual spaces, as illustrated in Fig. 2. This resulted in rich narratives where performers executed purposeful tasks, such as opening a drawer to retrieve an item, arranging objects on a desk, or preparing a simple meal, thereby ensuring that each motion served a clear function within a broader activity sequence. Further details of the *Scene-Driven LLM Scripting* process are provided in the supplemental materials (Fig. 14).

3.2. Environment and Capturing

Objects and Humans. HUMOTO is built on a carefully curated collection of 63 standard household objects, encompassing 72 distinct functional parts (Fig. 4). Unlike previous datasets relying on 3D scanning, we recruited professional

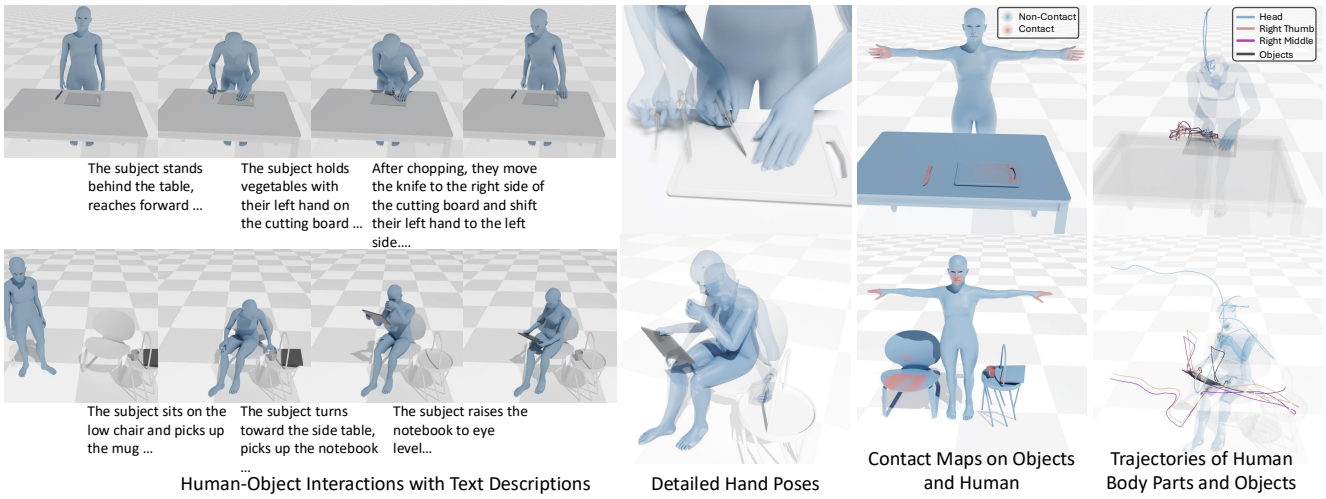


Figure 5. **HUMOTO dataset visualization.** We depict human-object interactions with text descriptions *(left)*, detailed hand poses, and contact maps highlighting interaction areas *(middle)*, and trajectories of human body parts and objects during activities *(right)*. These complementary representations provide comprehensive data for various applications.

artists to create precise digital models capturing crucial details, including articulated components and graspable surfaces. This ensures geometric accuracy while preserving part-level information essential for realistic interaction.

Our performer, outfitted with Rokoko smart-suits [53] and paired gloves, enabled high-fidelity tracking of full-body movements and finger articulations at 30 frames per second. The inertial sensor network provided reliable skeletal tracking, while the specialized gloves captured the fine-grained hand movements essential for natural object manipulation. The skeletal motions are transferred to a neutral human model with the standard Mixamo skeleton rigging [29].

Environment Setup. To minimize magnetic interference between the Rokoko suit's inertial sensors and metallic structures in the vicinity, we built a customized wooden stage to elevate performers from the floor by 12 inches. Two Kinect cameras were positioned at two corners, maximizing capture volume while minimizing occlusions during complex interactions. Spatial alignment between camera and motion capture systems was achieved using a calibration board to establish a common coordinate system. The dual-computer setup, one managing Kinect camera feeds and object tracking and the other handling Rokoko motion capture data, maintained precise temporal synchronization through UDP commands routed over a dedicated network.

Capturing process. We instructed performers to execute the scripted interactions with purpose rather than mechanical precision, maintaining the fidelity required for data analysis while preserving the characteristic fluidity of human motion. This approach was particularly important in capturing complex sequences with multiple objects, where performers might simultaneously engage with several items, *e.g.*, opening a drawer and reaching for an object, or repositioning multiple items on a surface. We captured these nuanced and complex multi-object interactions, including unconscious behaviors like fidgeting hands that characterize authentic human-environment engagement.

Processing the Raw Data. The technical processing pipeline addressed two primary challenges: temporal synchronization and spatial alignment between the human motion capture and object tracking data streams. At the beginning of each capture sequence, performers adopted a calibration stance at a predetermined position where the Rokoko system exhibited optimal tracking performance. This position, mapped to the camera coordinate system using our calibration board reference, established a transformation matrix that aligned both coordinate frames.

Object tracking leveraged the dual Kinect camera setup to minimize occlusions. The FoundationPose algorithm analyzed the visual data to determine 6DoF poses for each object. To address the limitations of frame-to-frame consistency assumptions during rapid movements, we implemented a dynamic reset mechanism based on mask pixel differences, reinitializing tracking when substantial movement was detected. To further improve the tracking result, we provide object masks by employing SAM2 with strategic human annotations, ensuring tracking consistency across frames where objects might be temporarily occluded.

3.3. Data Cleaning and Annotation

Multi-stage Quality Assurance. Our quality assurance protocol is a two-stage approach combining technical refinement and independent verification. During technical refinement, professional artists refined capture artifacts like drift and tracking errors, ensuring logical consistency in the interactions. During the subsequent verification, an independent team verified the sequences for natural and plausible human-object interactions, addressing issues such as

joint jitter and foot sliding. We iterated these two stages till all quality standards were met, ensuring fidelity to natural movements and interactions.

Textual Annotation. We invited an independent group to provide textual descriptions for each sequence based on the actual performance. These annotations included three elements: (1) a concise title highlighting the sequence's main goal with details on subtle differences, (2) a short script providing a complete yet brief description of the motion and interaction in the scene, and (3) a detailed long script elaborating on specific motions and interactions throughout the sequence. These multi-level textual annotations enhance the dataset's utility for applications requiring both visual and semantic understanding of human-object interactions.

This comprehensive approach with script generation, capture, processing, quality control, and annotation, resulted in a dataset that captures both the mechanics of object manipulations and the purposeful sequences in which these manipulations naturally occur. HUMOTO provides researchers with data reflecting how humans chain multiple actions together to achieve higher-level goals, enabling advances in human behavior prediction [27], robotic learning from demonstration [8, 13], virtual character animation [30, 38, 52], and augmented reality applications [36].

4. Dataset Analysis

This section elaborates on the quantitative and qualitative evaluations of the motion quality and compares HUMOTO against existing datasets: BEHAVE [1], OMOMO [39], IMHD [81], ParaHome [35] and GRAB [58].

4.1. Quantitative Evaluation

We evaluate human motion, object motion, and human-object interaction using metrics such as foot sliding, jerk, penetration, contact entropy, and state consistency. These metrics offer insights into motion quality, interaction realism, and the diversity of interaction states. Additionally, we introduce a new metric, **Motion Signal-to-Noise Ratio (MSNR)**, to assess the quality of motion relative to noise in the dataset. MSNR evaluates motion quality using the signal-to-noise ratio (SNR) [55] of joint kinematics. Higher SNR values indicate smoother motion, though excessive smoothing may result in loss of important details. We use Mixamo, an industry-standard motion capture dataset cleaned by artists, as the baseline for human motion quality. Datasets with MSNR values closer to Mixamo's indicate comparable motion quality. Further details and metric formulations are provided in Appendix A.2.1.

Comparison on Human and Object Motion. HUMOTO demonstrates superior performance in several key motion quality metrics. The data set exhibits the lowest foot sliding among all datasets compared, significantly outperforming established datasets like BEHAVE [1] and ParaHome [35]. This improvement can be attributed to our meticulous motion capture process and rigorous artist-led quality control. The low jerk values for human motion indicate smooth and natural movements, second only to IMHD [81], whose fast movements are more likely to have similar acceleration in a sequence. While the Mixamo dataset shows higher foot sliding and jerk, it is important to note that Mixamo contains specialized movements like street dancing, which inherently involves more dynamic motions that increase these metrics compared to typical HOI scenarios.

HUMOTO achieves 9.42 dB in Motion SNR, approaching Mixamo's reference value. This slightly lower SNR compared to IMHD and OMOMO stems from HUMOTO's complex interactions with detailed hand poses, which introduce higher frequency components often interpreted as "noise". Notably, OMOMO's combination of high SNR with high jerk values suggests clean signals that still contain abrupt motion changes, a phenomenon meriting future investigation. HUMOTO's high coherence demonstrates consistent, targeted motions while maintaining competitive diversity, especially compared to Mixamo. The unusually high diversity scores of other datasets may indicate excessive noise rather than true motion variety, artificially inflating their entropy measurements.

In object motion, HUMOTO demonstrates a notably low jerk, indicating realistic object manipulation, unlike the high values in OMOMO and IMHD. ParaHome's extremely low object jerk reflects that the objects in their long sequences are mostly static and barely interact with humans.

Comparison on Contact Quality. HUMOTO excels in contact quality metrics, achieving the lowest penetration among all datasets despite including detailed hand poses. This order of magnitude improvement over BEHAVE and OMOMO demonstrates our exceptional precision in capturing human-object spatial relationships, which is crucial for physically plausible interaction models. The contact entropy for HUMOTO shows a balanced distribution between contact states, more diverse than ParaHome but more focused than the potentially noisy patterns in IMHD and BEHAVE, suggesting meaningful interactions without excessive fluctuations. For state consistency, HUMOTO strikes a balance between the highly consistent but potentially oversimplified ParaHome and the less consistent BEHAVE, maintaining realistic transitions while avoiding rapid fluctuations that might indicate tracking errors.

Overall, HUMOTO combines the detailed hand articulation with superior metrics in foot sliding, smoothness of object motion, and minimal penetration, making it valuable for applications requiring physically accurate human-object interactions with natural motion.

Dataset	Key Statistics		Human Motion					Object Motion	Contact		
	Detailed Hand	Max. Objects in Scene	Foot Sliding (cm) ↓	Jerk (m/s^3) ↓	MSNR (dB) →	Coherence ↑	Diversity ↑	Jerk (m/s^3) ↓	Penetration (cm) ↓	Contact Entropy ↑	State Consistency ↑
BEHAVE [1]	✗	1	4.556	4.08	5.51	0.533	0.966	10.40	0.0606	2.2915	0.0667
OMOMO [39]	✗	1	2.130	15.10	12.37	0.619	0.978	27.40	0.0602	1.9468	0.4837
IMHD [81]	✓	1	1.474	**1.14**	14.20	0.554	0.951	24.06	0.1172	**2.4265**	0.2411
ParaHome [35]	✓	22	3.008	9.19	1.82	0.592	**0.980**	**0.08**	0.2167	1.0254	**0.6815**
HUMOTO	✓	15	**0.958**	1.87	**9.42**	**0.653**	0.956	1.13	**0.0068**	1.4587	0.5061
Mixamo	✓	-	3.184	8.14	10.88	0.616	0.958	-	-	-	-

Table 1. **Quantitative evaluation across human-object interaction datasets.** Metrics defined in Appendix A.2.1 should be interpreted holistically, as optimal values depend on specific applications. The table includes two additional statistical indicators that provide context for understanding dataset characteristics. Bold indicates **best**, underline indicates second-best. ↑: higher values are better, ↓: lower values are better, and →: values closer to Mixamo are better.

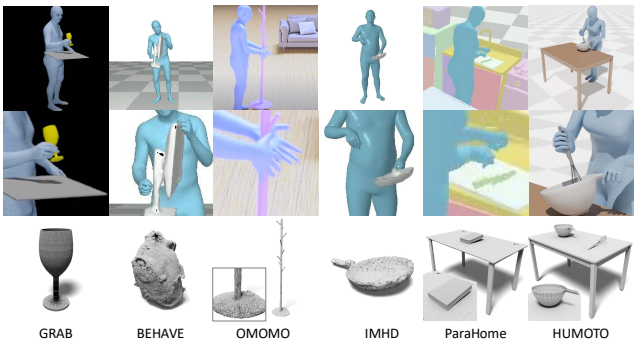


Figure 6. **Quality comparison.** We compare different datasets on motion dynamics, hand pose accuracy, and object meshes.

4.2. Qualitative Evaluation

The quantitative results are influenced by features of the datasets that do not necessarily represent quality issues. Therefore, they should be interpreted holistically rather than in isolation, as their values are influenced by multiple factors, including motion and interaction complexity. Thus, we also provide qualitative evaluations.

4.2.1. Visual Quality

We present a visual quality comparison in Fig. 6. While BEHAVE and OMOMO use only standard hand templates without detailed finger poses, IMHD offers finer hand modeling but exhibits significant penetration in several scenes. ParaHome provides relatively flexible hand motion, though their capture method (attaching tags on hands) interferes with natural movement, resulting in frequent clenched hand poses throughout the dataset. HUMOTO demonstrates superior hand pose quality, particularly during interactions. We also compare object mesh quality across datasets. Objects from prior datasets show noise artifacts due to 3D scanning limitations, while our object modeling pipeline produces clean, accurate representations.

4.2.2. Perceptual Study

To complement our quantitative analysis, we conducted a human perceptual study evaluating HUMOTO against existing HOI datasets through absolute quality assessment and direct pairwise comparison. We report the results of an online study taken by 26 participants, comprising students and researchers specializing in computational human motion.

Absolute Quality Assessment. Participants rated randomly selected videos from HUMOTO, BEHAVE, IMHD, OMOMO, and ParaHome on a 5-point Likert scale. HUMOTO achieved the highest scores in all categories: human motion (4.79±0.49), with 82% giving maximum scores), object motion (4.88±0.36), interaction quality (4.75±0.57), and overall quality (4.78±0.43). These scores significantly outperformed all comparison datasets, with the most notable difference in interaction quality, where BEHAVE scored only 2.48±1.05 and even recent datasets like IMHD (3.94±1.04) lagged considerably.

Pairwise Comparison. In this study, participants directly compared HUMOTO against other datasets showing the same interaction tasks. The results strongly favored HUMOTO in all dimensions, with 96% preferring HUMOTO over BEHAVE for overall quality. Even against newer datasets, HUMOTO was consistently preferred: 46% versus IMHD (with 50% rating both equally good), 65% versus OMOMO (28% ties), and 82% versus ParaHome (15% ties). For interaction quality specifically, HUMOTO outperformed BEHAVE (94% preference), OMOMO (65%), and ParaHome (67%), while against IMHD, HUMOTO was preferred by 38% and rated equally good by 46%.

These results demonstrate the superior quality of HUMOTO in both absolute ratings and direct comparisons, particularly for interaction quality and overall performance. Details are provided in Appendix A.3.

5. Discussions

Building upon the novel script generation pipeline, the multi-sensors motion capture system, and the rigorous quality control described in Sec. 3, the HUMOTO dataset provides not merely the detailed mechanics of object manipulation but the purposeful sequences in which these manipulations naturally occur. HUMOTO offers exceptional value

Figure 7. **Motion Generation by MotionGPT [30].** *Left:* Motion generated from the short scrip. *Mid:* Motion generated from the long script. *Right:* Motion with same text annotation from HUMOTO dataset.

for a wide range of research and applications, of which we highlight some below.

Human-Object Interaction and Motion Generation. Our dataset supports the development of generative models that can translate textual descriptions (*e.g.*, "pick up the coffee mug and drink from it") into realistic interaction sequences. The diversity of objects and interactions in HUMOTO provides rich supervision for text-conditioned motion synthesis. Our dataset is challenging as state-of-the-art human-object interaction models do not have the ability to generate interaction motion on multiple objects. To show this, we test MotionGPT [30] with HUMOTO prompts in Fig. 7. It appears that the model can generate a few reasonable motions based on the more abstract description, but fails to faithfully generate more fine-grained motions compared to the captured ground truth HUMOTO motions. This experiment demonstrates that state-of-the-art motion generation methods, despite being trained with large-scale datasets such as AMASS[47] and HumanML3D [21], still struggle with generating detailed human-object interaction. HUMOTO is designed to fill this gap.

Robotics and Embodied AI. The precision and diversity of interactions in HUMOTO make it particularly valuable for robotics research. To demonstrate the capability of our data, we use PyBullet [6] to compare HUMOTO with Parahome [35] in simulation settings. After weighting our objects and assigning similar mass to the ParaHome dataset, we use CoACD [65] to obtain convex shapes for simulation. Overlaying the final frame on the first (Fig. 8, Top) reveals significantly smaller object displacement in our dataset compared to Parahome, where interacted objects show substantial movement. Grasp synthesis, a popular robotic research topic [46, 62, 63], usually relies on simulated data that, despite passing simulator validation, often produces unnatural (*e.g.*, blue hand on mug bottom) or functionally unreasonable grasps (*e.g.*, fingers inside bowls). Comparing similar object grasps from HUMOTO with those from DexGraspNet [63] in Fig. 8 (Bottom) shows that our hand poses are more natural and aligned

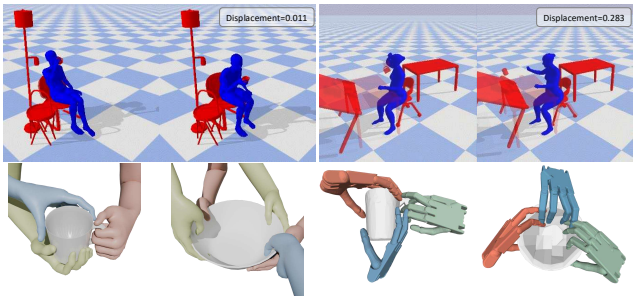


Figure 8. **Data for Robotics.** *Top:* Two simulator visualizations showing human sitting and holding mug. HUMOTO *(left)* displays minimal displacement, while ParaHome *(right)* shows significant object displacement during identical actions. *Bottom:* Hand manipulation comparison between HUMOTO *(left)* and simulated robotic grasps from DexGraspNet *(right)*.

with daily usage. Additionally, HUMOTO's task-oriented motion data can help robot learning systems develop capabilities directly transferable to real-world household assistance scenarios rather than simple interaction primitives.

Pose Estimation in Challenging Scenarios. State-of-the-art human pose estimation methods continue to face challenges in complex interaction scenarios. HUMOTO provides precise ground truth for these difficult cases with detailed hand articulation, particularly where objects partially occlude parts of the human body, creating ideal training data for models that must infer joint positions despite visual obstruction. Fig. 9 demonstrates how even the leading motion and pose estimation models, 4D Humans [19] and TRAM [64] struggle to predict correct poses from the renderings of our dataset. Additionally, none of these methods incorporates the hand pose estimation capabilities.

Authorized 2D Generation. Generating realistic images and videos often requires data that are difficult to capture, such as different viewpoints, object manipulation, or lighting changes. HUMOTO provides rich, human-involved scene data to simulate object addition/removal, reveal occluded areas, and capture lighting and shadow effects (Fig. 10, right). Existing 2D models, like Affordance Diffusion [74], often produce artifacts such as distorted hands and blurry poses (Fig. 10, left bottom). HUMOTO offers high-quality, realistic renderings of complex human-object interactions, enabling more accurate training for human-object interaction models [73, 74].

6. Conclusion and Limitations

In this work, we present HUMOTO, a comprehensive dataset of human-object interactions with detailed and accu-

Figure 9. **Human motion and pose estimation results on HUMOTO.** Comparison between 4D Humans [19] *(Mid)* and TRAM [64] *(Bottom)* on rendered images, showing estimated meshes (colored) against ground truth skeleton (white).

Figure 10. **Image editing.** *Left:* Hand-object interaction image generation conditioned on a mug. Recent work Affordance Diffusion *(bottom row)* produces physically implausible interactions with imprecise hand positioning, while HUMOTO can provide renderings *(top row)* of realistic hand placements at various positions. *Right:* Our dataset can also be used to provide renderings of object addition and removal, capturing differences in shadows and reflections, and facilitating authorized human-in-scene generative model training.

rate hand motion, and a dedicated scene-driven LLM scripting method to hierarchically design interaction scripts.

Despite HUMOTO's advancements, it has some limitations. First, due to motion capture suit size constraints, our dataset includes only a single performer, which may introduce a bias toward a particular human body shape and movement style. Second, the dataset preparation process required considerable manual cleaning and refinement of

the captured motion data. While such manual intervention ensures high-quality data, it represents a significant resource investment. To mitigate this challenge in future work, more advanced and robust pose estimation methods are needed. We hope that HUMOTO can serve as a foundational training set for developing such automated techniques, ultimately reducing the manual effort required for high-fidelity human-object interaction data collection.

Acknowledgement

The authors would like to thank Jimei Yang for their initial discussions on script development and Nathan Carr for designing and constructing the wooden stage for motion capture. We also thank Ziwen Chen for testing the MoCap suit, and Yuan Yao, Hanwen Jiang, Sanghyun Son, Shoubin Yu for helping us installing the objects. Qixing Huang is supported by NSF Career IIS-2047677, NSF IIS-2413161, and Gifts from Adobe and Google.

References

[1] Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022. 2, 3, 5, 6, 14, 16, 17

[2] Yukang Cao, Liang Pan, Kai Han, Kwan-Yee K. Wong, and Ziwei Liu. AvatarGO: Zero-shot 4d human-object interaction generation and animation. In *The Thirteenth International Conference on Learning Representations*, 2025. 1

[3] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:172–186, 2018. 2

[4] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023. 1

[5] Stefano Corazza, Lars Mündermann, Emiliano Gambaretto, Giancarlo Ferrigno, and Thomas P. Andriacchi. Markerless motion capture through visual hull, articulated icp and subject specific model generation. *International Journal of Computer Vision*, 87:156–169, 2010. 2

[6] Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning, 2016. 7

[7] Edilson de Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. Performance capture from sparse multi-view video. *ACM Trans. Graph.*, 27(3):1–10, 2008. 2

[8] Joseph DelPreto, Jeffrey Ian Lipton, Lindsay M. Sanneman, Aidan J. Fay, Christopher K. Fourie, Changhyun Choi, and Daniela Rus. Helping robots learn: A human-robot master-apprentice model using demonstrations via virtual reality teleoperation. *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10226–10233, 2020. 5

[9] Joseph DelPreto, Chao Liu, Yiyue Luo, Michael Foshey, Yunzhu Li, Antonio Torralba, Wojciech Matusik, and Daniela Rus. Actionsense: A multimodal dataset and recording framework for human activities using wearable sensors in a kitchen environment. In *Neural Information Processing Systems*, 2022. 2

[10] Christian Diller and Angela Dai. Cg-hoi: Contact-guided 3d human-object interaction generation. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2024. 1, 2

[11] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 14

[12] Adam Fishman, Adithyavairavan Murali, Clemens Eppner, Bryan Peele, Byron Boots, and Dieter Fox. Motion policy networks. In *Proceedings of the 6th Conference on Robot Learning (CoRL)*, 2022. 1

[13] Zipeng Fu, Tony Zhao, and Chelsea Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. *ArXiv*, abs/2401.02117, 2024. 2, 5

[14] Yasutaka Furukawa and Jean Ponce. Dense 3d motion capture from synchronized video streams. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. 2

[15] Juergen Gall, Carsten Stoll, Edilson de Aguiar, Christian Theobalt, Bodo Rosenhahn, and Hans-Peter Seidel. Motion capture using joint skeleton tracking and surface estimation. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1746–1753, 2009. 2

[16] Jiawei Gao, Ziqin Wang, Zeqi Xiao, Jingbo Wang, Tai Wang, Jinkun Cao, Xiaolin Hu, Si Liu, Jifeng Dai, and Jiangmiao Pang. CooHOI: Learning cooperative human-object interaction with manipulated object dynamics. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1

[17] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 409–419, 2017. 2

[18] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 409–419, 2017. 14

[19] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 7, 8

[20] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. *2018*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7297–7306, 2018. 2

[21] 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 (CVPR)*, pages 5152–5161, 2022. 7

[22] Vladimir Guzov, Julian Chibane, Riccardo Marin, Yannan He, Torsten Sattler, and Gerard Pons-Moll. Interaction replica: Tracking human–object interaction and scene changes from human motion. *2024 International Conference on 3D Vision (3DV)*, pages 1006–1016, 2022. 2

[23] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3193–3203, 2019. 2

[24] Shangchen Han, Beibei Liu, Robert Y. Wang, Yuting Ye, Christopher D. Twigg, and Kenrick Kin. Online optical marker-based hand tracking with deep labels. *ACM Transactions on Graphics (TOG)*, 37:1 – 10, 2018. 2

[25] Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. Synthesizing physical character-scene interactions. In *ACM SIGGRAPH 2023 Conference Proceedings*, New York, NY, USA, 2023. Association for Computing Machinery. 1, 2

[26] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8153–8163, 2024. 1

[27] Zhiming Hu, Zheming Yin, Daniel Haeufle, Syn Schmitt, and Andreas Bulling. Hoimotion: Forecasting human motion during human-object interactions using egocentric 3d object bounding boxes. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 5

[28] Yinghao Huang, Omid Tehari, Michael J. Black, Dimitrios Tzionas Max Planck Institute for Intelligent Systems, Tubingen, Germany, University of Amsterdam, Amsterdam, and The Netherlands. Intercap: Joint markerless 3d tracking of humans and objects in interaction. In *German Conference on Pattern Recognition*, 2022. 14

[29] Adobe Systems Inc. Mixamo, 2018. Accessed: 2025-03-07. 3, 4

[30] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36, 2024. 5, 7, 19

[31] Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Zhiyuan Zhang, Yixin Chen, Heng Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9331–9342, 2022. 14

[32] 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*, New York, NY, USA, 2024. Association for Computing Machinery. 1

[33] 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*, pages 1737–1747, 2024. 3, 14

[34] Roland Kehl and Luc Van Gool. Markerless tracking of complex human motions from multiple views. *Comput. Vis. Image Underst.*, 104:190–209, 2006. 2

[35] Jeonghwan Kim, Jisoo Kim, Jeonghyeon Na, and Hanbyul Joo. Parahome: Parameterizing everyday home activities towards 3d generative modeling of human-object interactions. *ArXiv*, abs/2401.10232, 2024. 1, 2, 3, 5, 6, 7, 14, 16, 17

[36] Yeonjoon Kim, Hangi Park, Seungbae Bang, and Sung-Hee Lee. Retargeting human-object interaction to virtual avatars. *IEEE Transactions on Visualization and Computer Graphics*, 22:2405–2412, 2016. 2, 5

[37] Makito Kobayashi, Chen-Chieh Liao, Keito Inoue, Sentaro Yojima, and Masafumi Takahashi. Motion capture dataset for practical use of ai-based motion editing and stylization, 2023. 2

[38] Lucas Kovar, Michael Gleicher, and Frédéric H. Pighin. Motion graphs. *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2002. 5

[39] Jiaman Li, Jiajun Wu, and C. Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42:1 – 11, 2023. 2, 3, 5, 6, 14, 16, 17

[40] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C. Karen Liu. Controllable human-object interaction synthesis. In *ECCV*, 2024. 1

[41] Liulei Li, Wenguan Wang, and Yi Yang. Human-object interaction detection collaborated with large relation-driven diffusion models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1, 2

[42] Jiehong Lin, Lihua Liu, Dekun Lu, and Kui Jia. Sam-6d: Segment anything model meets zero-shot 6d object pose estimation. *arXiv preprint arXiv:2311.15707*, 2023. 2

[43] Yunze Liu, Yun Liu, Chen Jiang, Zhoujie Fu, Kangbo Lyu, Weikang Wan, Hao Shen, Bo-Hua Liang, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20981–20990, 2022. 14

[44] Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding. *arXiv preprint arXiv:2401.08399*, 2024. 2, 3

[45] Matthew M. Loper, Naureen Mahmood, and Michael J. Black. MoSh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 33(6):220:1–220:13, 2014. 2

[46] Jiaxin Lu, Hao Kang, Haoxiang Li, Bo Liu, Yiding Yang, Qixing Huang, and Gang Hua. Ugg: Unified generative grasping. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXVII*, page 414–433, Berlin, Heidelberg, 2024. Springer-Verlag. 1, 7

[47] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2, 7

[48] Esteve Valls Mascaro, Daniel Sliwowski, and Dongheui Lee. HOI4ABOT: Human-object interaction anticipation for human intention reading collaborative roBOTs. In *7th Annual Conference on Robot Learning*, 2023. 1, 2

[49] Hyeongjin Nam, Daniel Sungho Jung, Gyeongsik Moon, and Kyoung Mu Lee. Joint reconstruction of 3d human and object via contact-based refinement transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10218–10227, 2024. 1

[50] Jeeseung Park, Jin-Woo Park, and Jong-Seok Lee. Viplo: Vision transformer based pose-conditioned self-loop graph for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17152–17162, 2023. 1, 2

[51] Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. *arXiv preprint arXiv:2312.06553*, 2023. 1, 2

[52] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Trans. Graph.*, 37(4):143:1–143:14, 2018. 1, 2, 5

[53] Rokoko. Smartsuit pro ii. https://www.rokoko.com/products/smartsuit-pro, 2022. 4

[54] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21064–21074, 2022. 2

[55] Claude E. Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37:10–21, 1949. 5

[56] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *ACM Trans. Graph.*, 38(6), 2019. 1

[57] Carsten Stoll, Nils Hasler, Juergen Gall, Hans-Peter Seidel, and Christian Theobalt. Fast articulated motion tracking using a sums of gaussians body model. *2011 International Conference on Computer Vision*, pages 951–958, 2011. 2

[58] 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 (ECCV)*, 2020. 2, 3, 5, 14

[59] Rong Tian, Zijing Zhao, Weijie Liu, Haoyan Liu, Weiquan Mao, Zhe Zhao, and Kimmo Yan. Samp: A model inference toolkit of post-training quantization for text processing via self-adaptive mixed-precision. In *Conference on Empirical Methods in Natural Language Processing*, 2022. 2

[60] Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stan Birchfield. Deep object pose estimation for semantic robotic grasping of household objects. *ArXiv*, abs/1809.10790, 2018. 2

[61] Weilin Wan, Lei Yang, Lingjie Liu, Zhuoying Zhang, Ruixing Jia, Yi-King Choi, Jia Pan, Christian Theobalt, Taku Komura, and Wenping Wang. Learn to predict how humans manipulate large-sized objects from interactive motions. *IEEE Robotics and Automation Letters*, 7(2):4702–4709, 2022. 2

[62] Weikang Wan, Haoran Geng, Yun Liu, Zikang Shan, Yaodong Yang, Li Yi, and He Wang. Unidexgrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3891–3902, 2023. 1, 7

[63] Ruicheng Wang, Jialiang Zhang, Jiayi Chen, Yinzhen Xu, Puhao Li, Tengyu Liu, and He Wang. Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11359–11366, 2022. 7

[64] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. In *European Conference on Computer Vision*, 2024. 7, 8

[65] Xinyue Wei, Minghua Liu, Zhan Ling, and Hao Su. Approximate convex decomposition for 3d meshes with collision-aware concavity and tree search. *ACM Transactions on Graphics (TOG)*, 41(4):1–18, 2022. 7

[66] 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 (CVPR)*, pages 17868–17879, 2024. 2

[67] Zhen Wu, Jiaman Li, Pei Xu, and C. Karen Liu. Human-object interaction from human-level instructions, 2024. 1

[68] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *ArXiv*, abs/1711.00199, 2017. 2

[69] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 1, 2

[70] Xianghui Xie, Xi Wang, Nikos Athanasiou, Bharat Lal Bhatnagar, Chun-Hao P. Huang, Kaichun Mo, Hao Chen, Xia Jia, Zerui Zhang, Liangxian Cui, Xiao Lin, Bingqiao Qian, Jie Xiao, Wenfei Yang, Hyeongjin Nam, Daniel Sungho Jung, Kihoon Kim, Kyoung Mu Lee, Otmar Hilliges, and Gerard Pons-Moll. Rhobin challenge: Reconstruction of human object interaction, 2024. 1

[71] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. InterDiff: Generating 3d human-object interactions with physics-informed diffusion. In *ICCV*, 2023. 1, 2

[72] Sirui Xu, Ziyin Wang, Yu-Xiong Wang, and Liangyan Gui. Interdreamer: Zero-shot text to 3d dynamic human-object interaction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1

[73] Zihui Xue, Mi Luo, Changan Chen, and Kristen Grauman. HOI-swap: Swapping objects in videos with hand-object interaction awareness. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 7

[74] Yufei Ye, Xueting Li, Abhinav Gupta, Shalini De Mello, Stan Birchfield, Jiaming Song, Shubham Tulsiani, and Sifei Liu. Affordance diffusion: Synthesizing hand-object interactions. In *CVPR*, 2023. 7

[75] Hangjie Yuan, Shiwei Zhang, Xiang Wang, Samuel Albanie, Yining Pan, Tao Feng, Jianwen Jiang, Dong Ni, Yingya Zhang, and Deli Zhao. Rlipv2: Fast scaling of relational language-image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21649–21661, 2023. 1

[76] Frederic Z Zhang, Yuhui Yuan, Dylan Campbell, Zhuoyao Zhong, and Stephen Gould. Exploring predicate visual context in detecting of human-object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10411–10421, 2023. 1

[77] Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Hoi-m3: Capture multiple humans and objects interaction within contextual environment. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 516–526, 2024. 2, 3

[78] Jason Y. Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2

[79] Xiaohan Zhang, Bharat Lal Bhatnagar, Vladimir Guzov, Sebastian Starke, and Gerard Pons-Moll. Couch: Towards controllable human-chair interactions. *ArXiv*, abs/2205.00541, 2022. 2

[80] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: Towards controllable human-chair interactions. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 2

[81] 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 (CVPR)*, pages 729–741, 2024. 2, 5, 6, 16, 17

[82] Yue Zhu, Nermin Samet, and David Picard. H3wb: Human3.6m 3d wholebody dataset and benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20166–20177, 2023. 3

A. Dataset Analysis Details

A.1. Statistics

We provide a visualization of statistics of the objects that occur in our dataset and the sequence duration distribution in Fig. 11. A detailed comparison of actions, objects and scenes between existing dataset statistics is included in the supp. material.

A.2. Metrics

A.2.1. Metrics Overview

To quantitatively assess the quality of our HUMOTO dataset compares to others, we define the following metrics that capture different aspects of motion naturalness and interaction accuracy.

For human and object motion: **Foot sliding** measures unnatural horizontal movement during ground contact. For foot joints below a height threshold, we calculate horizontal displacement with a weighting function that decreases as joints lift from the ground. Lower values typically indicate more natural motion. **Jerk** quantifies motion smoothness by measuring the rate of change of acceleration. Lower jerk represents smoother motions. **Motion Signal-to-Noise Ratio (MSNR)** evaluates motion quality through the SNR of joint kinematics. Higher SNR indicates smoother motion, though overly smoothed signals may lose important details. **Coherence** quantifies motion consistency by measuring pose cluster compactness. Values approaching 1 indicate highly consistent movement patterns with minimal deviation. **Diversity** measures variety of motion patterns using normalized Shannon entropy across pose clusters. Higher values indicate a wider range of motion patterns, though this may potentially identify jitter as diversity.

For interaction quality: **Penetration** assesses the physical plausibility of human-object interactions by measuring object intrusion into the human mesh. Lower values indicate more physically plausible interactions. **Contact entropy** quantifies the diversity of interaction states and transitions. Higher values indicate more diverse and complex interactions with a balanced distribution of contact behaviors. **State consistency** measures the temporal stability of interaction states, rewarding smooth contacts while penalizing rapid fluctuations. Higher scores indicate more consistent interaction states with fewer changes.

Jerk is computed for both human and object motion. Foot sliding, MSNR, Coherence, and Diversity apply only to human motion. Penetration, contact entropy, and state consistency evaluate human-object interaction quality. These metrics are influenced by features of the dataset that do not necessarily represent quality issues. Therefore, they should be interpreted holistically rather than in isolation, as their values are influenced by multiple factors including motion and interaction complexity. A complete definition of metrics is provided in Appendix A.2.2.

A.2.2. Metrics Formulation

Foot sliding measures unnatural horizontal movement during ground contact. For each foot joint j (ankles and toes) with height below threshold H_j, we compute:

$$\text{Sliding}_j = N_f \sum_{t \in \mathcal{S}_j} \|\mathbf{p}_{j,t+1}^{xy} - \mathbf{p}_{j,t}^{xy}\|_2 \cdot (2 - 2^{(\mathbf{p}_{j,t}^{z}/H_j)}) \quad (1)$$

where $\mathbf{p}_{j,t}$ is the position of joint j at frame t, $\mathcal{S}_j$ are frames where $\mathbf{p}_{j,t}^{z} < H_j$, and N_f is the total frame count. The exponential weighting function gradually decreases influence as joints lift from the ground. The final metric averages across all four foot joints and is reported in centimeters. In a standard setting, the lower the foot sliding value, the more natural the motion.

Jerk quantifies motion smoothness by measuring the rate of change of acceleration. For a sequence of joint positions $\mathbf{p}$ with N_f frames, we compute:

$$\text{Jerk} = \frac{1}{N_f - 3} \sum_{t=1}^{N_f - 3} \|\mathbf{a}_{t+1} - \mathbf{a}_t\|_2, \quad (2)$$

where velocities and accelerations are calculated as finite differences. As indicated here, lower jerk represents more smooth motions.

Motion Signal-to-Noise Ratio (MSNR) quantifies motion quality through the SNR of joint kinematics, computed as:

$$\text{SNR} = 10 \log_{10} \left(\frac{P_{\text{signal}}}{P_{\text{noise}}} \right) = 10 \log_{10} \left(\frac{\mathbb{E}[\hat{v}^2]}{\mathbb{E}[|v - \hat{v}|^2]} \right), \quad (3)$$

where v represents the normalized local joint velocities, and $\hat{v}$ is the temporally smoothed version of v obtained through convolution with a kernel size of 3. This metric captures the relationship between meaningful motion patterns and undesirable jitter or noise. A higher SNR value indicates a smoother motion. However, we should note that an overly smoothed signal may lose important details or contain less informative action.

Coherence score quantifies motion consistency by measuring pose cluster compactness. We compute coherence as

$$C = 1 - \frac{\mu_d}{\max_d}, \quad (4)$$

where μ_d is the mean distance from poses to their cluster centroids, and $\max_d$ is the maximum observed distance. Values approaching 1 indicate highly consistent movement patterns with minimal deviation.

Diversity metrics, on the other hand, quantify the variety of motion patterns in a dataset. We compute motion diversity using normalized Shannon entropy across pose clusters.

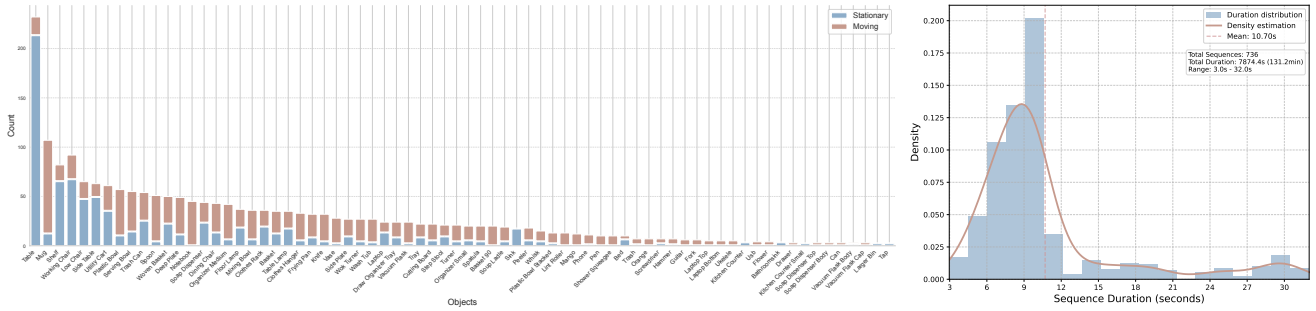


Figure 11. **Dataset statistics.** *Left:* Object occurrence frequency by motion type (stationary vs. moving). *Right:* Sequence duration distribution across the dataset.

Dataset	# hours	# subjects	# objects	# hands	Detailed Hand	body	Max. Obj. in Scene	setup
GRAB [58]	3.8	10	51	2	✓	✓	1	standing
BEHAVE [1]	4.2	8	20	-	✗	✓	1	portable
InterCap [28]	0.6	10	10	2	✓	✓	1	portable
OMOMO [39]	10.1	17	15	-	✗	✓	1	portable
FHPA [18]	0.9	6	26	1	✓	✗	1	room
HOI4D [43]	22.2	9	800	1	✓	✗	1	room
Chairs [31]	16.2	46	70	2	✓	✓	1	standing
ARCTIC [11]	1.2	10	11	2	✓	✓	1	standing
NeuralDome	4.6	10	23	2	✓	✓	1	standing
TRUMANS [33]	15	7	20	2	✓	✓	- (proxies)	room
ParaHome [35]	8.1	38	22	2	✓	✓	22	room
HUMOTO	2.2	1	63	2	✓	✓	15	scene

Table 2. **Dataset statistics (contd. from Fig. 11).** We provide details on the total durations, number of subjects, objects, presence of hand and body data, maximum objects in scene, and data collection setup styles.

After k-means clustering, we calculate

$$D = -\frac{\sum_{i=1}^{n} p_i \log_2 p_i}{\log_2 n}, \tag{5}$$

where p_i represents the proportion of frames in the i-th cluster. Higher diversity values indicate a wider range of motion patterns. However, this metric also identifies jittering or noise as diverse patterns.

Penetration quantifies the physical plausibility of human-object interactions by measuring object intrusion into the human mesh. For each frame, we sample points $\mathcal{P}_{obj}$ on object surfaces and compute the maximum penetration depth as:

$$\text{Penetration}(t) = \min_{p \in \mathcal{P}_{obj}} d(p, \mathcal{M}_h), \tag{6}$$

where $d(p, \mathcal{M}_h)$ is the signed distance from point p to the human mesh $\mathcal{M}_h$. Positive distances indicate interior points, with more positive values representing deeper penetration. We report the average maximum penetration across all frames, with lower values indicating more physically plausible interactions.

Contact entropy quantifies the diversity of interaction states and transitions during human-object interaction. For a sequence of interaction states discretized into categories (large penetration, contact, proximity, and distance), we compute:

$$\text{Entropy} = -\sum_{i,j} p(s_i \to s_j) \log_2 p(s_i \to s_j), \tag{7}$$

where $p(s_i \to s_j)$ is the probability of transitioning from state s_i to state s_j across all sampled points and frames. Higher entropy values indicate more diverse and complex interactions, with a balanced distribution of different types of contact and approach behaviors.

State consistency measures the temporal stability of interaction states, rewarding smooth and persistent contacts while penalizing rapid state fluctuations. For each sampled point, we calculate the average run length normalized by sequence length:

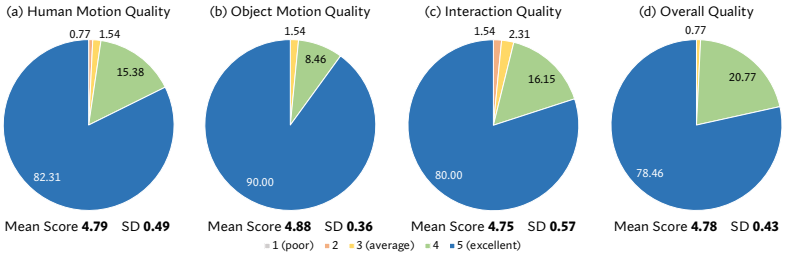
$$\text{Consistency} = \frac{1}{N_p} \sum_{p=1}^{N_p} \frac{\text{Avg. Run Length}_p}{\text{Sequence Length}}. \tag{8}$$

We additionally penalize points with large penetrations by

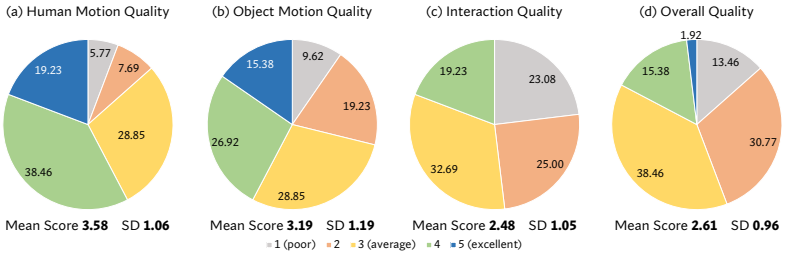
applying a scaling factor based on large penetration duration. Higher consistency scores indicate a more consistent interaction state with fewer state changes.

A.3. Perceptual Evaluation Results

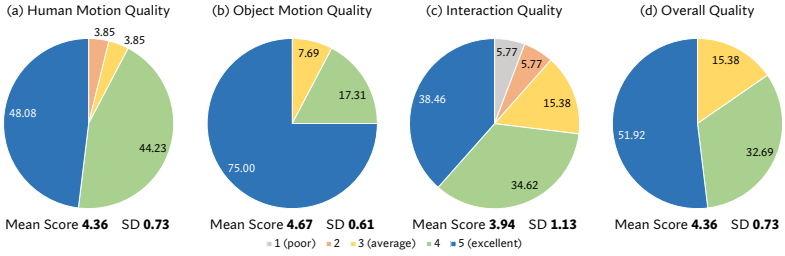
Following the details of our perceptual study setup provided in Sec. 4.2.2, we provide the detailed score distribution percentages of absolute quality evaluations in Fig. 12 and pairwise evaluations in Fig. 13.



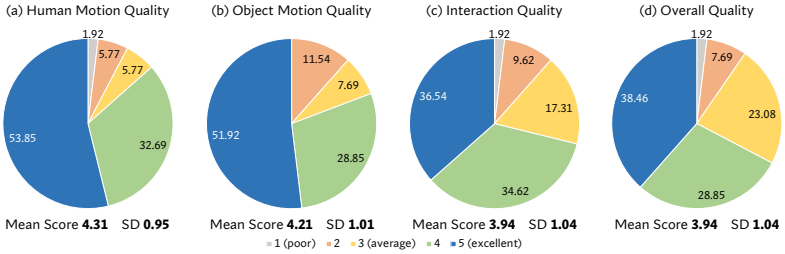
(a) HUMOTO



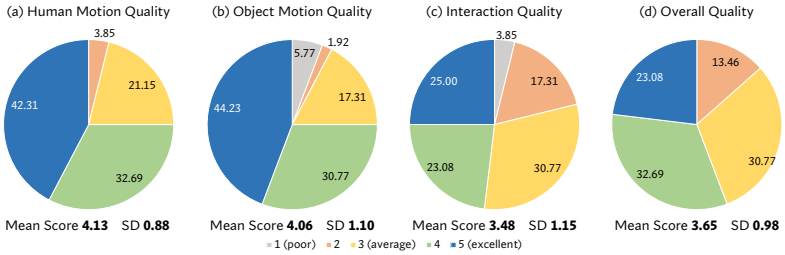
(b) BEHAVE



(c) OMOMO



(d) IMHD



(e) ParaHome

Figure 12. **Perceptual absolute quality ratings.** We show the aggregate percentages of absolute quality ratings on five-point Likert scales from our participants for HUMOTO, BEHAVE [1], OMOMO [39], IMHD [81], and ParaHome [35]. We assess the quality on four aspects: *(a) Human Motion Quality*, how plausible the human motions appear; *(b) Object Motion Quality*, how plausible the object motions appear; *(c) Interaction Quality*, how realistic the interactions between the humans and the objects appear; and *(d) Overall Quality*, how realistic the overall animations appear. We observe significant increases in ratings of 5 for HUMOTO in all four aspects.

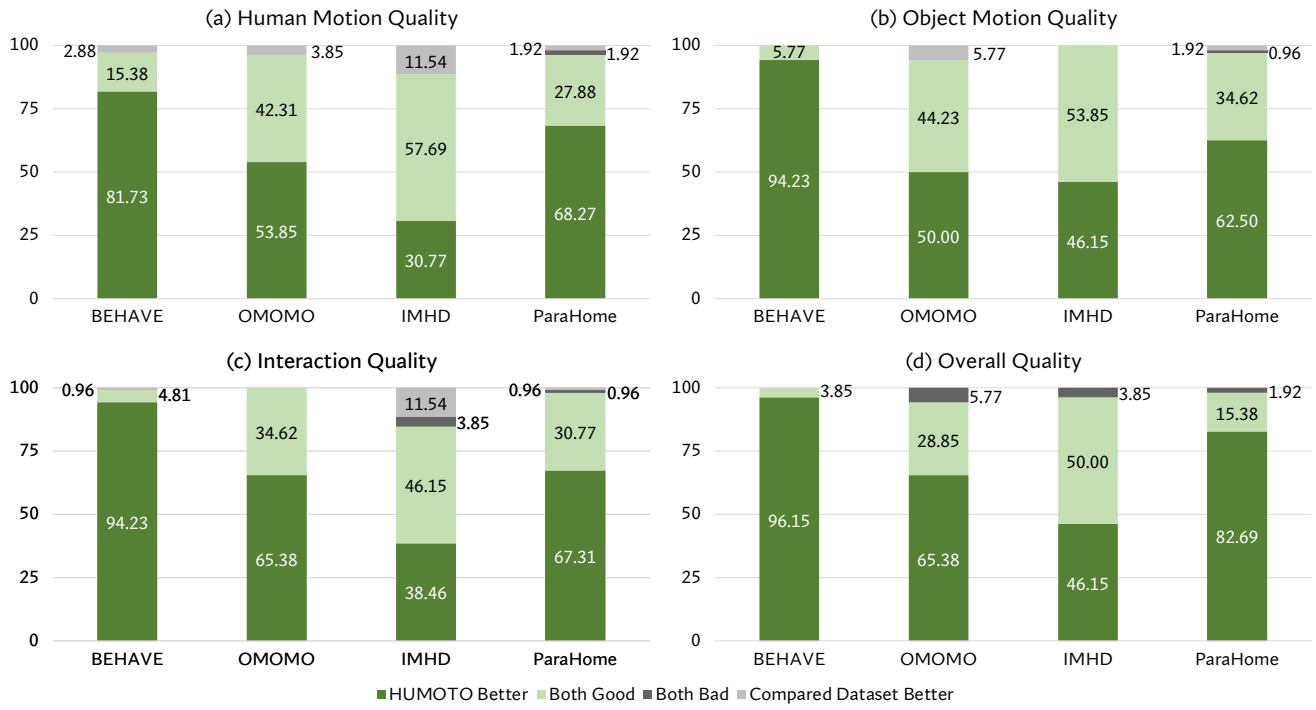


Figure 13. **Perceptual pairwise comparisons.** We show the aggregate percentages of pairwise comparison results from our participants, comparing side-by-side between HUMOTO and other datasets, including BEHAVE [1], OMOMO [39], IMHD [81], and ParaHome [35]. We assess the comparisons on four aspects: *(a) Human Motion Quality*, how plausible the human motions appear; *(b) Object Motion Quality*, how plausible the object motions appear; *(c) Interaction Quality*, how realistic the interactions between the humans and the objects appear; and *(d) Overall Quality*, how realistic the overall animations appear. After accounting for ties, we observe significant preferences for HUMOTO in all four aspects.

As an expert screenwriter and director, your task is to create scripts for a human-object interaction dataset. You'll be given a set of objects and a specific setting. Please follow these guidelines to create the script:
1.Object usage: Only use objects from the provided list. You may use all or a subset of these objects in each script. You cannot use any objects that not in the provided list.
2.Interaction focus: Each script should involve one person interacting with one or more objects.
3.Return format: Please return JSON output with format {"scene": <your_scene>, "short_script": <your_short_script>, "long_script": <your_long_script>, "objects": ["objects", "you", "used"]}.
Step 1: Set the scene: Envision a general setting where the given objects would naturally be found or follow the user provided setting.
Step 2: Write a concise script: Describe the interaction in 1-2 sentences, mentioning the objects involved.
Step 3: Write a longer script: Extend the interaction into a long paragraph. The goal is to create more accurate and detailed instructions that enable the actor to perform the intended actions precisely.
Step 4: List all objects you used in the scene, even those not provided originally.
setting: Your script should be based on a scene in a bedroom. You may select any combination of the provided objects. Your script should describe interaction lasts around 30 seconds.
objects: bed, clothes_hanger, clothes_rack, side_table, floor_lamp, mug, trash, trash_can, lint_roller, talbe_lamp, lower_chair.


```
{
  "scene": "Morning Outfit Selection",
  "short_script": "A person wakes up in bed, turns off the table lamp, walks to the clothes rack, and selects two clothes hangers with different outfits, holding each in front of themselves, examining their reflection in the mirror.",
  "long_script": "A person lying in bed stirs awake and slowly sits up, stretching and yawning. They reach over to the table lamp on the side table and touch the top of the lamp to turn it on. After a moment, they stand up from the bed and walk across the room toward a clothes rack filled with hanging garments. They pause in front of the rack, scanning the options, then carefully select a clothes hanger from the left side, lifting it off the rack with their right hand. They hold the hanger at arm's length, examining the garment. The person then walks to a mirror while holding the first hanger, positions themselves centered in front of the mirror, and holds the clothes on the hanger up against their body, tilting their head slightly while evaluating their appearance.",
  "objects": ["bed", "table_lamp", " side_table ", "clothes_rack", "clothes_hanger"]
}
```


As an expert screenwriter and director, your task is to create scripts for a human-object interaction dataset. Please follow the requirements and steps to write a script of human-object interaction based on the user given objects and its description:
1.Object usage: Only use the one provided object in your script.
2.Interaction focus: The interaction could only include one object and one performer. Your scripts should describe interactions lasts around 10 seconds.
3. Return format: Please return JSON output with format {"scene": <your_scene>, "short_script": <your_short_script>, "long_script": <your_long_script>, "objects": ["objects", "you", "used"]}.

<steps are the same as the scene-based scripts>

objects: mug: A ceramic mug with a comfortable handle, perfect for enjoying hot beverages like coffee or tea.


```
{
  "scene": "Mug Examination and Drinking",
  "short_script": "A person holds a mug, examines it, takes a sips, and sets it down again.",
  "long_script": "A person holds a ceramic mug by its handle, with their index finger threaded through the loop and their thumb resting on top. They lift the mug upward with a smooth motion, bringing it closer to their face and slightly rotating it to examine its contents. They then raise the mug to their lips, tilting it at approximately 30 degrees, and take a small sip. Finally, they lower the mug down to their waist level.",
  "objects": ["mug"]
}
```


Figure 14. **Examples of how we use LLMs to develop our human-object interaction scripts for capturing.** *Top:* We cluster objects into different scene types and create possible interactions within that scene. *Bottom:* For each individual object, we prompt LLMs on how one person would be possible to interact with the object.

Short script: The subject scoops ingredients using the spoon with left the hand from the deep plate. The subject adds ingredients with the left hand using the spoon to the mixing bowl. The subject mixes the content of the mixing bowl with the left hand.
Long script: The subject stand at the back of the table. The subject scoops ingredients inside deep plate with left hand using the spoon. The subject adds ingredients with left using spoon hand from deep plate to mixing bowl. The subject lifts the mixing bowl with the right hand. The subject inserts left hand into the mixing bowl. The subject mixes content with left hand inside the mixing bowl with.

Figure 15. **Motion generation results comparing our text-annotated dataset with MotionGPT [30].** *Top:* Generated motion sequence from short script input. *Middle:* Generated motion sequence from detailed long script input. *Bottom:* Ground truth motion sequence from our HUMOTO dataset. While MotionGPT can generate basic movements following general instructions, it struggles with the fine-grained hand-object interactions and precise manipulation sequences present in our dataset.