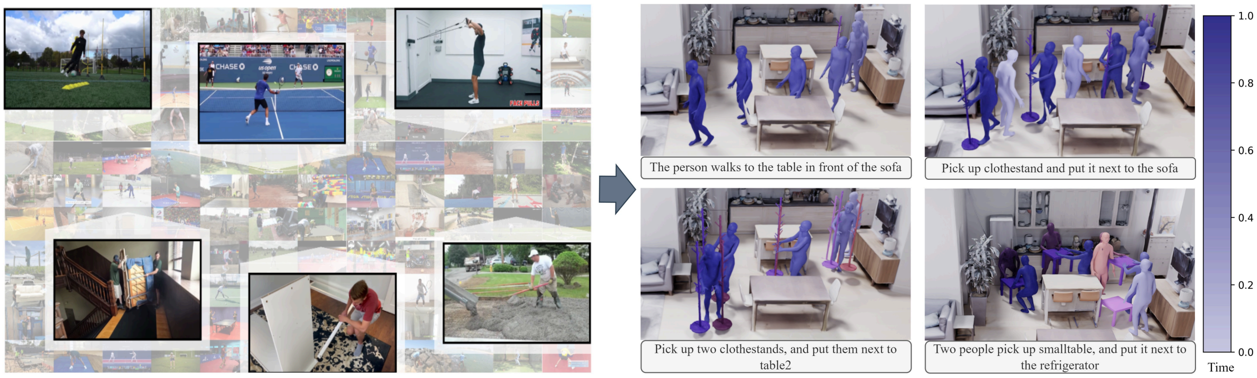
InterPose: Learning to Generate Human-Object Interactions from Large-Scale Web Videos

Yangsong Zhang[1], Abdul Ahad Butt[1], Gül Varol[2], Ivan Laptev[1]

[1]Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI)
[2]LIGM, École des Ponts, IP Paris, Univ Gustave Eiffel, CNRS
https://mael-zys.github.io/InterPose/


The person walks to the table in front of the sofa
Pick up clothestand and put it next to the sofa
Pick up two clothestands, and put them next to table2
Two people pick up smalltable, and put it next to the refrigerator
1.0
0.8
0.6
0.4
0.2
0.0
Time

(a) Sample videos with Human-Object Interactions (HOI)

(b) Zero-shot generation of interactions by our HOI-Agent

Figure 1. **Learning from online videos enables generation of complex human motions.** (a) Our InterPose dataset is obtained from videos with varying scenes and activities as well as diverse human-object interactions. (b) Our HOI-Agent deploys InterPose for training and enables zero-shot generation of collision-free navigation, human-object interactions and multi-person collaboration in 3D scenes.

Abstract

Human motion generation has shown great advances powered by recent diffusion models and large-scale motion capture data. Most of existing works, however, currently target animation of isolated people in empty scenes. Meanwhile, synthesizing realistic human–object interactions in complex 3D scenes remains a critical challenge in computer graphics and robotics. One obstacle towards generating versatile high-fidelity human-object interactions is the lack of large-scale datasets with diverse object manipulations. Indeed, existing motion capture data is typically restricted to single people and manipulations of limited sets of objects. To address this issue, we propose an automatic motion extraction pipeline and use it to collect interaction-rich human motions. Our new dataset InterPose contains 73.8K sequences of 3D human motions and corresponding text captions automatically obtained from 45.8K videos with human-object interactions. We perform extensive experiments and demonstrate InterPose to bring significant improvements to state-of-the-art methods for human motion generation. Moreover, using InterPose we develop an LLM-based agent enabling zero-shot animation of people interacting with diverse objects and scenes.

1. Introduction

Realistic animation of people in complex 3D environments remains a major challenge in gaming, robotics, virtual reality, and embodied AI. While games become more realistic, animation of interactions among characters and of object manipulations remains limited. In robotics, the deployment of complex manipulation skills is rare and often involves manual teleoperation. This motivates the growing interest in automatic generation of realistic interactions aiming to simulate continuous human behavior in 3D scenes.

Large-scale human motion datasets [15, 27, 35] have substantially advanced motion generation conditioned on past trajectories [61, 64], action categories [14, 41] and textual descriptions [15, 16, 42, 49, 62]. In contrast, Human–Object Interaction (HOI) [12, 25, 26, 37, 40], which requires generating coherent human–object motion with physically plausible contacts, remains a more challenging and insufficiently explored problem. Early efforts such as GOAL [47] and SAGA [52] primarily address small-object grasping, whereas subsequent methods [11, 25, 40, 55] extend to interactions with larger objects. However, these approaches typically depend on the paired human–object training data, exhibit limited generalization, and are often tailored for specific interaction scenarios. This limitation stems from the current HOI datasets [5, 26, 34, 46],

Dataset	Clip	Frames	Avg Frames	Avg Duration (s)	Hour
HD-VILA-100M [60]	10,042	1.6M	168.15	5.72	15.96
Kinetics-700 [7]	39,464	7.5M	190.21	6.40	70.20
Charades [45]	6,974	2.1M	304.24	11.53	22.34
Online videos	17,334	4.4M	255.59	8.36	40.24
Total	73,814	15.7M	213.34	7.25	148.74

Table 1. **Breakdown of InterPose video sources.** InterPose is obtained from online videos and three existing video datasets.

Dataset	Year	Source	Scale		Modality			Scene	
			Clip	Hour	Motion	Text	Interaction	Indoor	Outdoor
AMASS [35]	2019	C	11,265	40	B,H	✗	✗	✓	✗
HumanML3D [15]	2022	C	14,616	28.6	B	✓	✗	✓	✗
Motion-X [27]	2023	C,V	81,084	144.2	B,H	✓	✗	✓	✓
GRAB [46]	2020	C	1,334	3.8	B,H	✗	✓	✓	✗
BEHAVE [5]	2022	C	321	4.1	B	✗	✓	✓	✗
OMOMO [26]	2023	C	-	10.1	B	✓	✓	✓	✗
TRUMANS [19]	2024	C	-	15.0	B,H	✗	✓	✓	✗
LINGO [18]	2024	C	-	16.0	B,H	✓	✓	✓	✗
HIMO [34]	2024	C	3,376	9.4	B,H	✓	✓	✓	✗
InterAct-X [57]	2025	C	16,201	30.7	B,H	✓	✓	✓	✗
ParaHome [21]	2025	C	207	8.1	B,H	✓	✓	✓	✗
HUMOTO [31]	2025	C	735	2.2	B,H	✓	✓	✓	✗
InterPose	2025	V	73,814	148.7	B,H	✓	✓	✓	✓

Table 2. **Comparison with existing human motion datasets.** Letters C and V mean the data is obtained from motion capture and videos, respectively. B and H present body and hand motion, respectively. Note that, although InterAct-X merges and augments many existing HOI datasets, InterPose provides the largest-scale interaction data.

which remain constrained in both scale and diversity, covering only a narrow spectrum of interaction types.

To address these challenges, in this work we develop a pipeline for automatic collection of large-scale and diverse human motion data from videos. We build on advances in 3D human pose estimation [44] and vision–language models (VLMs) [4] and extract interaction-rich 3D human motions representing a large variety of daily living activities and sports. Using this pipeline, we introduce **InterPose**, a large-scale dataset of human motions and corresponding text descriptions, focused on diverse interactions obtained from YouTube videos and open-source video datasets (cf. Table 1). As shown in Table 2, InterPose is the largest interaction-focused public dataset with human motions. As opposed to motion capture datasets taken in controlled settings, InterPose also maximizes diversity due to ”in the whild” nature of source videos. We use our dataset to train human animation models MaskedMimic [48] and OmniControl [54], and demonstrate significant improvements in human motion brought by InterPose.

We leverage the scale and the rich diversity of InterPose and demonstrate zero-shot HOI generation using MaskedMimic [48]. Experiments on standard HOI benchmarks, including OMOMO [26] and BEHAVE [5], demonstrate that our zero-shot approach obtains excellent results and outperforms other in-domain trained methods such as CHOIS [25]. The benefit of InterPose is particularly prominent for the BEHAVE [5] benchmark, where all models are assessed in zero-shot settings. These results demonstrate that using InterPose improves generalization and enables zero-shot HOI generation for new objects.

To extend interaction generation to 3D environments, we further introduce HOI-Agent, a zero-shot LLM-based framework. HOI-Agent integrates a high-level LLM planner with a low-level motion generator (i.e., MaskedMimic [48]). The planner produces detailed execution steps, collision-free waypoints, and executable Python code for downstream control, while the InterPose-powered motion generator synthesizes human–object interactions in a zero-shot manner. As illustrated in Figure 1, HOI-Agent can handle a broad range of interactions.

In summary, our contributions are threefold:

- We develop an automatic pipeline for data collection and introduce InterPose, a large-scale 3D human motion dataset with diverse interactive activities, facilitating generation of human-object interactions.
- We use InterPose to train state-of-the-art spatial control methods and demonstrate substantial performance improvements enabled by our dataset.
- We present HOI-Agent, a unified zero-shot framework that integrates our motion model to perform diverse manipulation tasks and generalizes to novel scenarios.

2. Related work

We briefly review relevant works on controllable motion generation, 3D HOI datasets, as well as HOI generation.

Controllable human motion generation. Human motion generation has been explored under various conditions, including action classes [14, 41], textual descriptions [16, 42, 49, 62], speech audio [24, 66], music [22, 50], scene characteristics [51, 65] and objects [25, 26]. Recent approaches further focus on controllable motion generation methods with text descriptions, including temporal or spatial composition [2, 3, 43], style control [9, 17] and spatial guidance [10, 20, 48, 54]. To achieve zero-shot object interaction, we sample the contact points and then control the person's hands to move with the object. To this end, the spatial controllability is essential in our zero-shot framework.

Spatial control models such as OmniControl [54] and MaskedMimic [48] are all trained with human-only datasets [15, 35], whose data mainly contains locomotion and limited interactions. Such models, hence, exhibit difficulties in generating human-object interactions (HOI). To enable HOI generation, we retrain [48, 54] using InterPose and show improvements.

Human-object interaction generation. With the emergence of full-body motion datasets together with hand-object interactions [46], models have been developed to synthesize full-body motions leading up to object grasping [47, 52]. Subsequent works [6, 13, 19] have expanded this task to human-object interaction for more complex object manipulations. With more recent datasets [5, 26], several works further explore interaction with larger objects [11, 25, 40, 55]. Specifically, CHOIS [25] generates human-object interactions conditioned on sparse 2D object waypoints while HOIFHLI [53] further integrates an LLM-based agent to enable the interaction ability in 3D scenes. Previous kinematic-based methods may generate implausible motions such as sliding, floating and penetration. To solve these problems, physics-based methods have recently been explored [6, 12, 33, 37, 58]. InterMiMic [58] focuses on imitating the motion capture data in a physics simulator while correcting contact artifacts. Prior works highly rely on paired human and object data, and the performance is thus limited by the dataset scale. Zero-shot HOI methods have recently been explored [23, 30, 56]. Specifically, ZeroHSI [23] and ZeroHOI [30] both leverage a video generation model to first generate human-object videos and then extract the human and object

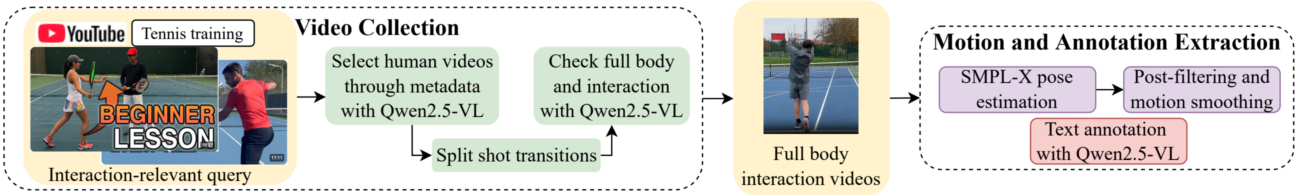

YouTube
Tennis training
BEGINNER LESSON
Interaction-relevant query
Video Collection
Select human videos through metadata with Qwen2.5-VL
Check full body and interaction with Qwen2.5-VL
Split shot transitions
Full body interaction videos
Motion and Annotation Extraction
SMPL-X pose estimation
Post-filtering and motion smoothing
Text annotation with Qwen2.5-VL

Figure 2. **Overview of data collection for the InterPose dataset.** Our framework contains a module for collecting interaction-rich videos (left) and a module for automatic extraction of 3D human motions and corresponding text captions (right).

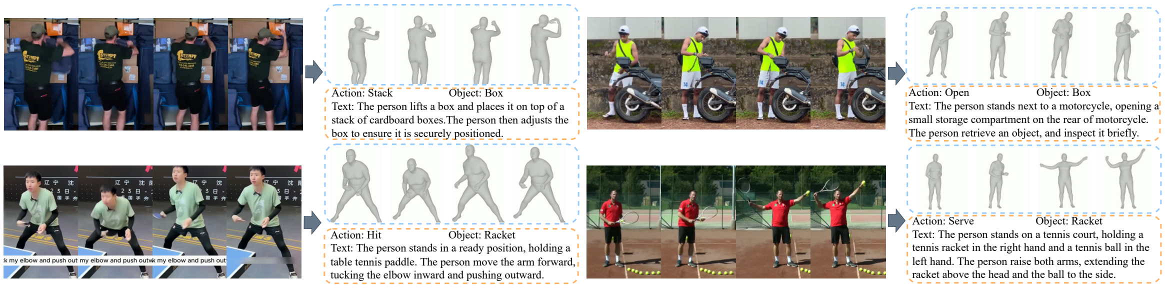

Action: Stack
Object: Box
Text: The person lifts a box and places it on top of a stack of cardboard boxes.The person then adjusts the box to ensure it is securely positioned.
Action: Open
Object: Box
Text: The person stands next to a motorcycle, opening a small storage compartment on the rear of motorcycle. The person retrieve an object, and inspect it briefly.
Action: Hit
Object: Racket
Text: The person stands in a ready position, holding a table tennis paddle. The person move the arm forward, tucking the elbow inward and pushing outward.
Action: Serve
Object: Racket
Text: The person stands on a tennis court, holding a tennis racket in the right hand and a tennis ball in the left hand. The person raise both arms, extending the racket above the head and the ball to the side.

Figure 3. **Examples from InterPose.** Our 3D human motion data originates from diverse interaction scenarios including working actions, sports activities, indoor and outdoor scenes. All motion sequences are annotated with action, object label and a detailed textual description.

motion through estimation and optimization. Despite the high interaction quality, the full process is time-consuming, which takes several hours to obtain one sequence. Concurrent work such as HOIGen [28] explores HOI video collection in the context of video generation, whereas our focus is on the 3D generation.

Due to data scarcity, existing research mainly focuses on certain types of interaction, and can hardly generalize to daily scenarios. Based on InterPose, we enable MaskedMimic [48] zero-shot HOI ability across various objects. Moreover, we explore an LLM agent framework to perform various HOI tasks such as navigation in 3D scenes and multi-person collaboration in zero-shot manner.

Human-object interaction datasets. The field of human motion generation has advanced rapidly with the availability of large motion capture datasets. AMASS [35] provides a comprehensive corpus of 3D human motion that has substantially facilitated motion modeling. Motion-X [27] extends this line of work by aggregating multiple existing datasets and extracting motion parameters from videos to form an even larger human-only collection. However, current large-scale datasets mainly include isolated human motions such as walking and sports activities, and contain only a limited set of sequences with interactions.

While there exist human-object datasets with static scene objects [36, 51, 63], we focus on HOI datasets with dynamic object manipulation. GRAB [46] proposes a dynamic HOI dataset where the interaction is mainly focused on small objects. More recent datasets [5, 26] have introduced human motion involving manipulation of a wider range of object sizes. Prior datasets focus only on single object interaction, to address this, HIMO [34] is proposed specifically for multi-object interaction. Due to the difficulty of collecting large-scale human-object motion, the existing HOI datasets are small-scale. Although InterAct-X [57] merges and augments previous HOI datasets, the total duration (31 hours) is still lower than AMASS [35] (40 hours). The lack of large and diverse datasets hinders the progress of realistic interaction synthesis. In contrast, as shown in Table 2, our proposed dataset, InterPose, contains large-scale interactive actions for both indoor and outdoor scenes, spanning 149 hours. We will make InterPose and derived models publicly available to foster progress in HOI generation.

3. Automatic dataset construction

In this paper, we propose a large-scale human motion dataset, **InterPose**, which mainly focuses on interactions. As illustrated in Figure 2, the motion data collection pipeline is composed of 7 main components: 1) designing keywords to query human-object interaction videos from the web; 2) checking video metadata such as thumnail and tags with a VLM before downloading; 3) pre-processing such as shot splitting 4) pre-filtering videos without full-body humans or without interactions; 5) whole-body 3D pose estimation; 6) post-processing through smoothing and filtering floating or static humans data; 7) textual annotation through a VLM. In Section 3.1, we present details on the collection of videos with interaction-rich human motion. Automatic human motion estimation and captioning steps are described in Section 3.2.

3.1. Video collection

As detailed in Table 1, we construct InterPose by processing 14K raw videos from the indoor action video dataset Charades [45], 51K from the human action recognition dataset Kinetics-700 [7], and 61K from the high-resolution general video dataset HD-VILA-100M [60], yielding 56K motion sequences after filtering and extraction. To further enrich data diversity, we additionally collect 29K online videos and extract 17K motion clips from YouTube. The pipeline for existing datasets is identical to that for online videos, except that the query step is omitted. Details of the online video collection are provided in the following paragraphs.

Interaction-relevant query design. Manually searching for HOI videos is both time-consuming and labor-intensive. To

Control joints	Training dataset	AMASS [35]			OMOMO [26]			BEHAVE[5]		
		Succ 0.5 ↑	Succ 0.2 ↑	MPJPE ↓	Succ 0.5 ↑	Succ 0.2 ↑	MPJPE ↓	Succ 0.5 ↑	Succ 0.2 ↑	MPJPE ↓
Pelvis	AMASS [35]	**0.9657**	0.8673	**46.1**	0.9834	0.9087	49.0	0.9932	0.9444	49.9
	OMOMO [26]	0.8146	0.5748	150.4	0.9585	0.7759	70.0	0.9824	0.8035	74.8
	InterPose	0.9228	0.8021	97.6	**0.9979**	0.9191	59.0	0.9892	0.9593	55.0
	InterPose + AMASS [35]	0.9636	**0.9135**	55.4	**0.9979**	**0.9855**	**41.2**	**0.9959**	**0.9905**	**37.1**
Hands	AMASS [35]	0.9451	0.7841	72.7	0.8382	0.5290	133.9	0.9837	0.7724	89.8
	OMOMO [26]	0.7145	0.2757	207.5	0.9170	0.6017	116.5	0.9688	0.6125	121.8
	InterPose	0.8831	0.8010	131.1	0.9191	0.8071	79.2	0.9919	0.9607	65.6
	InterPose + AMASS [35]	**0.9521**	**0.8717**	**71.5**	**0.9751**	**0.9004**	**60.9**	**0.9973**	**0.9837**	**52.5**

Table 3. **Evaluation of human motion controllability using physics-based MaskedMimic [48] generator.** We train MaskedMimic [48] with different datasets and then evaluate all the models with different control settings on the human-only dataset AMASS [35], and object interaction datasets (OMOMO [26] and BEHAVE [5]). The best results are in bold, and the second best results are underlined.

Control joints	Training dataset	HumanML3D [15]				OMOMO [26]				BEHAVE[5]			
		FS ↓	Traj err 0.5 ↓	Traj err 0.2 ↓	MPJPE ↓	FS ↓	Traj err 0.5 ↓	Traj err 0.2 ↓	MPJPE ↓	FS ↓	Traj err 0.5 ↓	Traj err 0.2 ↓	MPJPE ↓
Pelvis	HumanML3D [15]	0.0585	0.0957	0.3467	0.0708	0.0945	0.0604	0.5208	0.0836	0.0497	0.0179	0.3259	0.0668
	OMOMO [26]	0.0830	0.3125	0.5283	0.1786	0.1175	0.0646	0.4375	0.0719	0.0334	0.0045	0.2277	0.0492
	InterPose	**0.0539**	0.0957	0.3066	0.0708	**0.0779**	0.0208	0.2292	0.0515	**0.0320**	**0**	0.0982	0.0342
	InterPose + HumanML3D [15]	0.0656	**0.0508**	**0.2207**	**0.0489**	0.1185	**0.0042**	**0.1167**	**0.0327**	0.0488	**0**	**0.0402**	**0.0232**
Hands	HumanML3D [15]	0.0681	0.1377	0.5645	0.0849	0.0960	0.2313	0.8646	0.1275	0.0878	0.0670	0.6741	0.0911
	OMOMO [26]	0.1066	0.4736	0.8936	0.2016	0.1408	0.1187	0.8771	0.1410	0.0768	0.1696	0.7232	0.1223
	InterPose	**0.0436**	0.1699	0.6123	0.0950	**0.0596**	0.0979	0.6312	**0.0059**	**0.0275**	0.0580	0.3839	0.0591
	InterPose + HumanML3D [15]	0.0524	**0.0605**	**0.3320**	**0.0555**	0.0741	**0.0354**	**0.3396**	0.0461	0.0639	**0.0045**	**0.2098**	**0.0320**

Table 4. **Evaluation of human motion controllability using diffusion-based OmniControl [54] generator.** Similar to Table 3, we train OmniControl [54] with different datasets and then evaluate on HumanML3D [15], OMOMO [26] and BEHAVE [5].

address this issue, we leverage a large language model (LLM) to automatically generate a comprehensive set of human interaction queries, enabling systematic coverage of diverse interaction types. Specifically, we ask LLM to give all the queries which may possibly contain full-body and interaction videos such as tennis training, fencing and watering flowers. Besides, LLM also provides a variety of queries for similar topics, e.g., for tennis, LLM generates queries such as *tennis training*, *tennis tutorial*, *tennis match* and *tennis serve*. These queries are integrated into our fully-automated pipeline for searching and downloading online videos. As shown in Figure 3, the collected videos cover diverse indoor and outdoor scenarios, daily object manipulation actions and sports. More examples can be found in the supplementary material.

Metadata check. Building upon the query design stage, we incorporate an early-stage filtering mechanism to avoid unnecessary downloads. Specifically, we employ Qwen2.5-VL-7B [4] to analyze available metadata, including titles, categories, thumbnails, and ASR-derived speech transcripts, before downloading. Qwen2.5-VL is prompted to verify whether each candidate video predominantly features a full-body human and whether it likely involves interaction with objects or other individuals. This pre-selection significantly reduces the number of irrelevant videos.

Shot detection. Given the downloaded candidate videos, we apply PySceneDetect [8] to segment videos into coherent temporal clips. When splitting, the frames per second (FPS) of all segments are processed to 30 which is the most common FPS for training motion models. Short segments (i.e., <2 sec.) and low resolution (<360p.) videos are discarded in this step to maintain motion quality. Besides, long videos are also split into multiple clips of no more than 30 seconds, facilitating later processing and annotation while preserving motion diversity.

Full-body and interaction verification. As metadata-based filtering cannot guarantee accuracy, we perform a finer-grained validation on each retained segment. We use a pretrained 2D keypoint detector MediaPipe [32] to infer the keypoints of the most clearly visible human. We then filter the videos where humans are mostly occluded with average confidence less than 0.5 for key joints or the ratio of full-body frames lower than a predefined threshold. In addition, we prompt Qwen2.5-VL [4] again for all remaining video segments to further confirm that the human consistently engages in physical interactions with objects or other people across most frames. This two-step verification removes residual false positives and ensures the collected clips provide clear, high-quality interaction content.

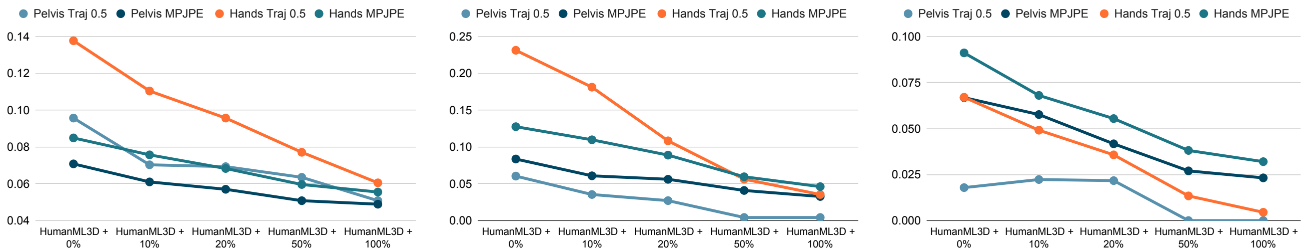
3.2. 3D motion and annotation extraction

Pose estimation. Our objective is to extract plausible body pose parameters (i.e., SMPL-X [38]) with both body and hand motions, and to obtain a text description for each interaction. In terms of body pose, we leverage WHAM [44] to jointly estimate the local human pose and the global transformations in world coordinates. As WHAM only estimates SMPL [29] parameters, we further apply HaMeR [39] to extract the detailed finger motion. To enhance the consistency between body and hands predictions, we input the same ViTPose++ [59] 2D pose predictions for WHAM [44] and HaMeR [39] pipelines. An additional optimization is applied to merge body and hand parameters to avoid twisted wrists. Note that we extract only the human motion, and not the object, mainly due to the limited performance of current 3D object reconstruction methods. We show that this is already a crucial step for obtaining large-scale human motion data with interaction priors. Future work can revisit complementing the data with object motion.

Post-filtering and motion smoothing. Human bounding boxes

Control joints	Filtering	HumanML3D [15]				OMOMO [26]				BEHAVE[5]			
		FS ↓	Traj err 0.5 ↓	Traj err 0.2 ↓	MPJPE ↓	FS ↓	Traj err 0.5 ↓	Traj err 0.2 ↓	MPJPE ↓	FS ↓	Traj err 0.5 ↓	Traj err 0.2 ↓	MPJPE ↓
Pelvis	✗	0.0713	0.0527	0.2441	0.0522	0.1215	**0**	0.1208	0.0346	0.0622	**0**	0.0622	0.0268
	✓	**0.0656**	**0.0508**	**0.2207**	**0.0489**	**0.1185**	0.0042	**0.1167**	**0.0327**	**0.0488**	**0**	**0.0402**	**0.0232**
Hands	✗	0.0683	0.0703	0.3584	0.0590	0.0791	**0.0333**	**0.3292**	0.0470	0.0684	0.0089	0.2188	0.0337
	✓	**0.0524**	**0.0605**	**0.3322**	**0.0555**	**0.0741**	0.0354	0.3396	**0.0461**	**0.0639**	**0.0045**	**0.2098**	**0.0320**

Table 5. **Ablation on data filtering with OmniControl [54].** We train OmniControl on the combination of HumanML3D [15] and InterPose with/without filtering. All the models are evaluated on HumanML3D [15], OMOMO [26] and BEHAVE [5] test sets with pelvis or hands control.



(a) Evaluation on HumanML3D [15] test set

(b) Evaluation on OMOMO [26] test set

(c) Evaluation on BEHAVE [5] test set

Figure 4. **Impact of dataset size.** We train OmniControl [54] on HumanML3D [15] dataset and subsets of InterPose of different sizes. The subsets are composed of 0%, 10%, 20%, 50%, and 100% of InterPose training set. All the models are evaluated on HumanML3D [15], OMOMO [26] and BEHAVE [5] test sets. We report Traj 0.5 and MPJPE on pelvis control and hands control setting for all the models.

Eval joints	Training joints	Full trajectory			Avg		
		Traj err 0.5 ↓	Traj err 0.2 ↓	MPJPE ↓	Traj err 0.5 ↓	Traj err 0.2 ↓	MPJPE ↓
Pelvis	One [54]	**0.0605**	**0.2637**	**0.0564**	**0.0404**	**0.1623**	**0.0367**
	Random	0.0957	0.3467	0.0708	0.0633	0.2182	0.0499
Hands	One [54]	0.5195	0.9014	0.2161	0.2482	0.4730	0.1102
	Random	**0.1377**	**0.5645**	**0.0849**	**0.0686**	**0.2639**	**0.0423**

Table 6. **Ablation for OmniControl [54] training setting on the HumanML3D [15] dataset.** We train OmniControl [54] with different number of joints and then evaluate with 2 settings on HumanML3D [15]. Avg means that we evaluate with 5 sparsity levels in the controlling signal, including 1, 2, 5, 25% density, and 100% density (full trajectory) and then report the average performance.

that are relatively small within a video make it difficult for pose estimation and VLM to capture accurate information. We therefore discard detected people with bounding box area below a threshold of 5,000p. We also detect and discard motions with abrupt translations or orientations, e.g., the motion may jump from one person to another because of the detection errors. Furthermore, the human actions are manually slowed down or even paused in some videos such as sports tutorial videos. We discard static data by setting a threshold on joint velocities.

Finally, due to object occlusions, or motion blur, the initial SMPL-X [38] results may exhibit jitter issues. To tackle jitter problems, we first detect frames with severe discontinuities in human motion and apply Slerp to interpolate motion trajectories. We then apply another smoothing step to all frames using a small window size to remove the small-scale jitter while preserving motion details.

Text annotation. In terms of textual motion descriptions, we carefully prompt a VLM, Qwen2.5-VL-7B [4], to automatically annotate each human in the candidate videos. As InterPose mainly focuses on interactive actions, we ask the VLM to give the whole-body action, a detailed text about the contact body part and the corresponding objects. In the multi-person video cases, to ensure the correspondence between the human and each text, we crop the video using the bounding box given from previous ViTPose++ [59] predictions. As illustrated in Figure 3, in addition to textual motion descriptions, the VLM is also used to annotate the human action and the interacted object. All the detailed prompts can be found in the supplementary material.

4. Experiments

In this section we present advances in human motion generation brought by InterPose. In particular, we deploy InterPose to train two state-of-the-art methods, namely OmniControl [54] and MaskedMimic [48], and demonstrate improvements on standard benchmarks for controlled motion generation in Section 4.1. InterPose analysis and ablation experiments are presented in Section 4.2. In Section 4.3 we evaluate InterPose benefits on the task of generating Human-Object Interactions (HOI). Finally, we present a HOI-Agent and demonstrate its zero-shot generalization to new objects and tasks in Section 4.4.

4.1. Controllable motion generation

Implementation details. To validate the advantage of InterPose, we train spatial control models (i.e., OmniControl [54] and MaskedMimic [48]) on different datasets: general human motion datasets (AMASS [35] and HumanML3D [15]) and interaction datasets (OMOMO [26] and InterPose). Since interactive tasks often involve multiple control joints, we modify the original training procedure of OmniControl [54], which samples only one joint per sample, by allowing a random number of joints to be controlled. This adjustment better reflects real application scenarios. An ablation study is provided in Section 4.2. Otherwise, the training settings follow those of OmniControl [54] and MaskedMimic [48]. To evaluate spatial controllability, we condition on full joint trajectories for motion capture datasets AMASS [35], HumanML3D [15], and OMOMO [26], as most interactions require consistent contact along the trajectory. Ad-

Model	Training datasets	Human motion		Interaction				
		$H_{feet} \downarrow$	FS $\downarrow$	$C_{prec} \uparrow$	$C_{rec} \uparrow$	$C_{f1} \uparrow$	$C_{percent}$	$P_{hand} \downarrow$
CHOIS [25]	OMOMO [26]	4.47	$\underline{0.3435}$	**0.7982**	0.6477	0.6773	0.5602	0.6135
MaskedMimic [48]	AMASS [35]	$\underline{0.15}$	**0.3317**	0.7882	0.6197	0.6484	0.6225	0.6565
	OMOMO [26]	0.76	0.4752	0.7798	0.6032	0.6335	0.5978	**0.5876**
	InterPose	0.47	0.4249	0.7890	$\underline{0.7185}$	$\underline{0.7172}$	$\underline{0.7134}$	$\underline{0.6131}$
	InterPose + AMASS [35]	**−0.11**	0.4823	$\underline{0.7915}$	**0.7597**	**0.7483**	**0.7464**	0.6193

Table 7. **Evaluation of human-object interaction generation on the OMOMO [26] dataset.** We train MaskedMimic [48] with different human motion datasets and then evaluate the object interaction ability by interpolating object motion and controlling human hands to follow the objects.

Model	Training datasets	Human motion		Interaction				
		$H_{feet} \downarrow$	FS $\downarrow$	$C_{prec} \uparrow$	$C_{rec} \uparrow$	$C_{f1} \uparrow$	$C_{percent}$	$P_{hand} \downarrow$
CHOIS [25]	OMOMO [26]	1.91	0.4546	0.6994	0.5813	0.5873	0.6296	**0.6889**
MaskedMimic [48]	AMASS [35]	-0.32	**0.3320**	**0.7020**	0.5737	0.5685	0.6454	0.8097
	OMOMO [26]	**−0.01**	0.4468	0.6975	0.5510	0.5575	0.6195	$\underline{0.8025}$
	InterPose	$\underline{0.03}$	$\underline{0.4460}$	0.6977	$\underline{0.6191}$	$\underline{0.5988}$	$\underline{0.6992}$	0.8098
	InterPose + AMASS [35]	-0.31	0.4904	$\underline{0.7015}$	**0.6936**	**0.6490**	**0.7706**	0.8887

Table 8. **Evaluation of human-object interaction generation on the BEHAVE [5] dataset.** Similar to Table 7, we report the performance of the models in Table 7 on the BEHAVE [5] dataset.

ditionally, we assess zero-shot generalization on BEHAVE [5] to examine performance under unseen control distributions.

Metrics. Following MaskedMimic [48] and OmniControl [54], the spatial controllability is evaluated by the success rate (Succ), failure rate (Traj err) and mean per-joint position error (MPJPE) to calculate the overall accuracy and position error on the control joints. For example, Succ 0.5 denotes the proportion of generated samples whose maximum keyframe position error is below 0.5m. In addition, we report the foot sliding (FS) metric for OmniControl [54] evaluation to assess motion realism.

Results. The performance of MaskedMimic [48] and OmniControl [54] trained on different datasets is presented in Tables 3-4 respectively. For MaskedMimic [48], training on AMASS [35] yields strong performance under pelvis control due to the diverse locomotion, which helps the model follow pelvis trajectories. However, controlling two hands is more challenging and closely tied to interaction, leading to a significant performance drop. In this setting, training on InterPose substantially outperforms AMASS [35], especially on the test sets of interaction datasets OMOMO [26] and BEHAVE [5]. Since InterPose contains no motions from these datasets, the improved generalization highlights the quality and diversity of its interactions. Combining AMASS and InterPose achieves the best overall performance. Similarly, for OmniControl [54], training with InterPose notably enhances hand controllability. As MaskedMimic [48] and OmniControl [54] represent fundamentally different paradigms, their consistent improvements demonstrate the robustness of the InterPose dataset.

Methods	Preference on OMOMO [26] (%)	Preference on BEHAVE [5] (%)
CHOIS [25]	38.5	23.5
Ours	**61.5**	**76.5**

Table 9. **User studies on OMOMO [26] and BEHAVE [5] datasets.** ”Ours” indicates trained on the combination of our InterPose and AMASS [35] datasets.

4.2. Ablations

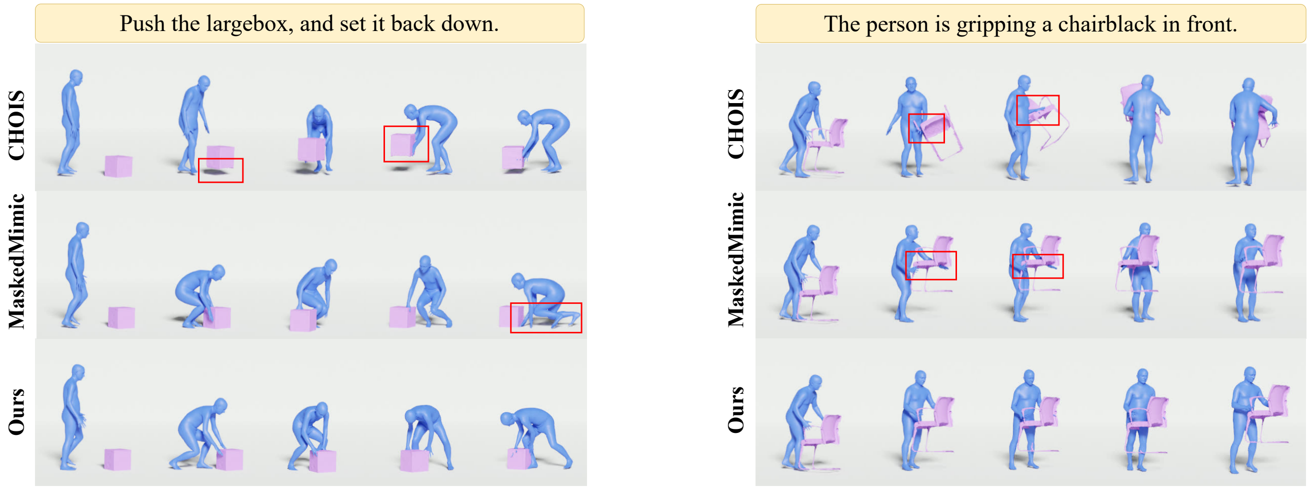
Impact of post-filtering. We evaluate the effect of post-filtering described in Section 3.2 by comparing the spatial controllability of OmniControl [54] trained on the combined HumanML3D and InterPose datasets, with and without post-filtering, as shown in Table 5. Model with post-filtering achieves better overall performance. Although the unfiltered model only performs slightly worse for controllability, the lack of filtering introduces abnormal artifacts such as floating and abrupt rotations, leading to significantly increased foot sliding. These results highlight the importance of post-filtering for improving motion realism.

Impact of dataset size. We investigate how dataset size influences performance to assess whether larger scale interaction data further enhances spatial control. OmniControl [54] is trained on HumanML3D [15] combined with varying proportions of InterPose data (0%, 10%, 20%, 50%, and 100%). As shown in Figure 4, spatial controllability is evaluated on HumanML3D [15], OMOMO [26], and BEHAVE [5] test sets. Performance improves consistently as the amount of training data increases across all control settings and benchmarks, suggesting that collecting additional motion data could yield further gains. Detailed quantitative results are provided in the supplementary material.

OmniControl training setting. We further investigate the impact of training configurations in OmniControl [54]. The original model is trained by randomly selecting a single control joint at each iteration. In contrast, we retrain the model by sampling a random number of control joints per iteration and compare its performance with the original setting. Evaluation is conducted under five levels of control signal sparsity, and we report both the full-trajectory results and average performance across sparsity levels. Note that, the Traj 0.5 and MPJPE for ”Avg“ column of OmniControl [54] pelvis control are from their published results. Otherwise, we report our reproduced numbers. As shown in Table 6, while our strategy shows a slight decrease in pelvis control accuracy, it consistently yields substantial improvements in hand control performance.

4.3. Application: Zero-shot HOI generation

Implementation details. To demonstrate the zero-shot HOI generation capability of InterPose, we conduct experiments on


Push the largebox, and set it back down.
CHOIS
MaskedMimic
Ours
The person is gripping a chairblack in front.
CHOIS
MaskedMimic
Ours

(a) Visual comparison on OMOMO [26] dataset.

(b) Visual comparison on BEHAVE [5] datasets.

Figure 5. **Visual comparison of CHOIS [25], MaskedMimic [48] and Ours when evaluated on OMOMO [26] and BEHAVE [5] datasets.** Note that original MaskedMimic [48] is trained on AMASS [35] dataset and Ours denotes MaskedMimic trained on the combination of InterPose and AMASS [35]. Red boxes highlight examples of unrealistic human action, unrealistic object motion or interaction.

the standard benchmark [25], where the model must generate human–object motion conditioned on sparse object waypoints. For this task, we adopt the spatially controllable network MaskedMimic [48]. However, since MaskedMimic [48] is primarily trained on human-only datasets [35] with limited interactive motion, its ability to follow object control points is constrained. To overcome this limitation, we retrain the model on InterPose and evaluate its performance on HOI generation tasks. We further assess human motion quality and interaction fidelity by comparing against the supervised HOI framework CHOIS [25] on OMOMO [26] and BEHAVE [5], using the same object waypoints for fair comparison. Object trajectories and rotations are interpolated using PCHIP and Slerp, respectively, and contact points are sampled based on object mesh points to guide hand control during motion generation. To identify hand-object contact points, we deploy a simple heuristic and sample points on the object surface that have smaller distance to human hands.

Metrics. For HOI generation, we follow CHOIS [25] and measure foot height (H_{feet}) and foot sliding (FS) to obtain the human motion quality. Contact precision (C_{prec}), recall (C_{rec}), F1 score (C_{f1}), contact percentage (C_{percent}) and penetration score (P_{hand}) are also calculated to evaluate the interaction performance, following prior work [25]. More details about the evaluation metrics are provided in Section A of the supplementary material.

Results. We compare our models with CHOIS [25] on the HOI benchmarks presented in Tables 7-8. MaskedMimic [48] trained on OMOMO [26] obtains the worst results, because OMOMO [26] is small-scale, and thus the model fails to generalize due to the gap between the GT and the interpolated hands' trajectory. While the model trained on the AMASS [35] dataset can achieve promising human motion quality, but due to lack of interactive training behaviors, the model fails to accurately follow the trajectories of interactive hands, resulting in suboptimal interaction performance. Incorporating InterPose data substantially enhances interaction quality, particularly in terms of object contact metrics. Notably, even though CHOIS [25] is trained in a fully-supervised manner on the OMOMO [26] dataset, models trained on InterPose still outperform CHOIS [25] in contact quality under zero-shot generation. Furthermore, in evaluations on the BEHAVE [5] dataset, where all models are assessed in zero-shot settings, our models consistently exceed CHOIS [25] across most metrics. The consistent performance across both OMOMO [26] and BEHAVE [5] demonstrates the robust zero-shot generalization capability of models trained on InterPose.

User study. We conduct a user study to further evaluate and compare the zero-shot HOI performance of our method with the supervised HOI approach CHOIS [25]. Specifically, we generate 20 HOI sequences using CHOIS [25] and MaskedMimic [48] trained on a combination of InterPose and AMASS [35] (here denoted as Ours) for both the OMOMO [26] and BEHAVE [5] datasets. Twenty participants were recruited to indicate their preference based on two criteria: (1) the quality of human and object motion, and (2) the accuracy of contact performance. As reported in Table 9, the consistent results across participants confirm the effectiveness and robustness of InterPose in facilitating high-quality interaction generation. Notably, the advantage becomes more evident on the out-of-distribution BEHAVE [5] dataset, which indicates the model trained on InterPose could exhibit superior generalization and robustness in zero-shot interaction scenarios.

Qualitative results. We qualitatively compare CHOIS [25] with zero-shot HOI generation using MaskedMimic [48], and MaskedMimic trained on the combination of InterPose and AMASS [35] (Ours). Evaluation is conducted on the HOI datasets OMOMO [26] and BEHAVE [5], as illustrated in Figure 5. From the example in Figure 5a, we observe that CHOIS [25] generates unrealistic object motion (the object appears to float even when the person is distant) and the motion is not consistent with the textual description (e.g., the person is lifting rather than pushing the object). For the original MaskedMimic [48], the person struggles to follow the object trajectory and nearly falls due to distributional shifts.

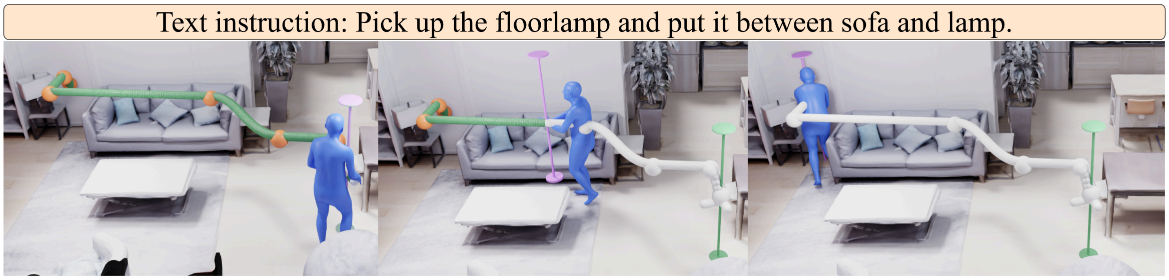

Text instruction: Pick up the floorlamp and put it between sofa and lamp.

Figure 6. **Illustration of zero-shot human and object motion generation by our HOI-Agent.** Given an initial scene and text instruction, the LLM planner generates a detailed plan and collision-free sparse waypoints (orange). The contact points sampled from the object surface (green) are then used as hand control points for human motion generation.

In Figure 5b, both CHOIS [25] and MaskedMimic [48] fail to produce plausible contacts: CHOIS [25] exhibits varying contact positions, whereas MaskedMimic [48] suffers from penetration issues. These examples demonstrate that a human motion generator trained on InterPose results in interaction motions of higher quality compared to other methods. More visual results are provided in the supplementary material.

4.4. Application: Interaction in 3D Scenes

Previous section addresses the task of HOI generation given object waypoints and no scene information. Here we extend this approach to fully-automatic HOI generation and address multiple interaction types including collision-free navigation, multi-person collaboration as well as single and multi-object manipulation. Some methods, such as CHOIS [25] and CooHOI [12], primarily focus on specific interaction types, whereas others, e.g., TokenHSI [37], support diverse interaction generation, but require manual task planning and transition handling, which is labor-intensive. To overcome these limitations, we formulate the problem as interaction generation within a given 3D environment guided by textual instructions. We introduce a LLM-based agent framework, **HOI-Agent**, which performs automatic task planning and enables zero-shot generation of diverse interaction behaviors in complex 3D scenes.

HOI-Agent framework. The HOI-Agent framework consists of a LLM-based high-level planner and a low-level motion generator. Given a human-level instruction and an environment state, including the position, orientation, and size of each object in the scene, the LLM planner analyzes spatial relationships and generates detailed step-wise plans along with collision-free human or object waypoints. As described in Section 4.3, the sparse waypoints are then interpolated into full object motion. Based on the object motion, the contact points are sampled from the object surface to guide a human motion generator. The low-level generator, MaskedMimic [48] trained on InterPose, supports zero-shot HOI generation and can be conditioned on the generated waypoints to execute the desired interactions. In this manner, HOI-Agent enables flexible and robust interaction generation across complex 3D scenes. More detailed description, prompts and an illustrative example are provided in the supplementary material.

Quantitative results. We evaluate our proposed HOI-Agent by generating 20 HOI sequences in a 3D scene. The success rate is computed based on two criteria: (1) the manipulated object reaches a position close to the target location, and (2) the human remains upright without falling. Among the 20 generated sequences, 13 are successful, while the remaining failures are attributed to either wrong planning (3 cases) or the human subject losing balance and falling (4 cases).

Qualitative results. As illustrated in Figure 6, the high-level planner is able to generate instruction-following and scene-consistent waypoints used to control human motion generation. We also show the capabilities of our HOI-Agent in Figure 1. With our proposed HOI-Agent framework, we enable the automatic motion generation for various tasks: collision-free human navigation, HOI in new 3D scenes, multi-object interaction and multi-person collaboration. More results can be found on the project page [1].

5. Conclusion

In this paper we proposed InterPose, a large-scale and automatically created dataset with 3D human motions containing diverse human-object interactions. InterPose enables both kinematics-based and physics-based models to generate higher quality motion sequences. We further explored zero-shot HOI generation, and demonstrated advantages of our dataset compared to previous training setups. Moreover, we proposed HOI-Agent enabling fully-automatic zero-shot HOI generation in unseen complex 3D scenes.

Limitations and future work. First, InterPose is currently composed of human motions only and excludes associated object movements. With the progress in 3D object reconstruction, future work can augment InterPose with the shape and motion of manipulated objects to provide additional data for HOI generation. Second, while our HOI-Agent enables high-level planning, collision-free trajectory generation, and interaction motion synthesis, its effectiveness depends on the underlying LLM and prompt engineering. Learning HOI-Agent could be another interesting direction for future work. Overall, we believe that InterPose and the proposed HOI-Agent framework will advance and inspire further research in HOI generation.

Acknowledgement. This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean Government (MSIT) (No. RS-2024-00457882, National AI Research Lab Project).

References

[1] InterPose project webpage. https://mael-zys.github.io/InterPose/. 8, 1

[2] Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. TEACH: Temporal Action Compositions for 3D Humans. In *International Conference on 3D Vision (3DV)*, 2022. 2

[3] Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. SINC: Spatial composition of 3D human motions for simultaneous action generation. In *ICCV*, 2023. 2

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv*, 2025. 2, 4, 5

[5] 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 *CVPR*, 2022. 1, 2, 3, 4, 5, 6, 7

[6] Jona Braun, Sammy Christen, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Physically plausible full-body hand-object interaction synthesis. In *3DV*, 2024. 2

[7] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv*, 2019. 2, 3

[8] Brandon Castellano. Pyscenedetect. https://github.com/Breakthrough/PySceneDetect, 2021. 4

[9] chuan guo, Yuxuan Mu, Xinxin Zuo, Peng Dai, Youliang Yan, Juwei Lu, and Li Cheng. Generative human motion stylization in latent space. In *ICLR*, 2024. 2

[10] Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. MotionLCM: Real-time controllable motion generation via latent consistency model. In *ECCV*, 2024. 2

[11] Christian Diller and Angela Dai. CG-HOI: Contact-guided 3D human-object interaction generation. In *CVPR*, 2024. 1, 2

[12] 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 *NeurIPS*, 2024. 1, 2, 8

[13] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. IMoS: Intent-driven full-body motion synthesis for human-object interactions. In *Computer Graphics Forum*, 2023. 2

[14] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2Motion: Conditioned generation of 3D human motions. In *ACM MM*, 2020. 1, 2

[15] 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 *CVPR*, 2022. 1, 2, 4, 5, 6

[16] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. MoMask: Generative masked modeling of 3D human motions. In *CVPR*, 2024. 1, 2

[17] Deok-Kyeong Jang, Soomin Park, and Sung-Hee Lee. Motion Puzzle: Arbitrary motion style transfer by body part. *ACM Transactions on Graphics (TOG)*, 2022. 2

[18] 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. 2

[19] 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 *CVPR*, 2024. 2

[20] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *ICCV*, 2023. 2

[21] Jeonghwan Kim, Jisoo Kim, Jeonghyeon Na, and Hanbyul Joo. ParaHome: Parameterizing everyday home activities towards 3D generative modeling of human-object interactions. In *CVPR*, 2025. 2

[22] Buyu Li, Yongchi Zhao, Shi Zhelun, and Lu Sheng. Dance-Former: Music conditioned 3D dance generation with parametric motion transformer. In *AAAI*, 2022. 2

[23] Hongjie Li, Hong-Xing Yu, Jiaman Li, and Jiajun Wu. ZeroHSI: Zero-shot 4D human-scene interaction by video generation. *arXiv*, 2024. 2

[24] Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. Audio2Gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *ICCV*, 2021. 2

[25] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object interaction synthesis. *arXiv*, 2023. 1, 2, 6, 7, 8

[26] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 2023. 1, 2, 3, 4, 5, 6, 7

[27] 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. *NeurIPS*, 2023. 1, 2, 3

[28] Kun Liu, Qi Liu, Xinchen Liu, Jie Li, Yongdong Zhang, Jiebo Luo, Xiaodong He, and Wu Liu. HOIGen-1M: A large-scale dataset for human-object interaction video generation. In *CVPR*, 2025. 3

[29] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 2015. 4

[30] Yuke Lou, Yiming Wang, Zhen Wu, Rui Zhao, Wenjia Wang, Mingyi Shi, and Taku Komura. Zero-shot human-object interaction synthesis with multimodal priors, 2025. 2

[31] Jiaxin Lu, Chun-Hao Paul Huang, Uttaran Bhattacharya, Qixing Huang, and Yi Zhou. HUMOTO: A 4D dataset of mocap human object interactions. *arXiv*, 2025. 2

[32] Cameron Lugaresi, Jiuqiang Tang, Hartwig Nash, Chris McClanahan, Mark Zhu, Chuo-Ling Chang, Ming Guang Yong, Joe Lee, Fan Pang, Victor Fu, et al. MediaPipe: A framework for building perception pipelines. *arXiv*, 2019. 4

[33] Zhengyi Luo, Jinkun Cao, Sammy Christen, Alexander Winkler, Kris Kitani, and Weipeng Xu. Grasping diverse objects with simulated humanoids. *arXiv*, 2024. 2

[34] Xintao Lv, Liang Xu, Yichao Yan, Xin Jin, Congsheng Xu, Shuwen Wu, Yifan Liu, Lincheng Li, Mengxiao Bi, Wenjun Zeng, and Xiaokang Yang. HIMO: A new benchmark for full-body human interacting with multiple objects. In *ECCV*, 2024. 1, 2, 3

[35] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. AMASS: Archive of motion capture as surface shapes. In *ICCV*, 2019. 1, 2, 3, 4, 5, 6, 7

[36] Zoltán Á Milacski, Koichiro Niinuma, Ryosuke Kawamura, Fernando de la Torre, and László A Jeni. GHOST: Grounded human motion generation with open vocabulary scene-and-text contexts. *arXiv*, 2024. 3

[37] Liang Pan, Zeshi Yang, Zhiyang Dou, Wenjia Wang, Buzhen Huang, Bo Dai, Taku Komura, and Jingbo Wang. TokenHSI: Unified synthesis of physical human-scene interactions through task tokenization. In *CVPR*, 2025. 1, 2, 8

[38] 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 *CVPR*, 2019. 4, 5

[39] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024. 4

[40] 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*, 2023. 1, 2

[41] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In *ICCV*, 2021. 1, 2

[42] Mathis Petrovich, Michael J. Black, and Gul Varol. TEMOS: Generating diverse human motions from textual descriptions. In *ECCV*, 2022. 1, 2

[43] Mathis Petrovich, Or Litany, Umar Iqbal, Michael J Black, Gül Varol, Xue Bin Peng, and Davis Rempe. Multi-track timeline control for text-driven 3D human motion generation. *CVPRW*, 2024. 2

[44] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. WHAM: Reconstructing world-grounded humans with accurate 3D motion. In *CVPR*, 2024. 2, 4

[45] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016. 2, 3

[46] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *ECCV*, 2020. 1, 2, 3

[47] Omid Taheri, Vasileios Choutas, Michael J Black, and Dimitrios Tzionas. GOAL: Generating 4D whole-body motion for hand-object grasping. In *CVPR*, 2022. 1, 2

[48] Chen Tessler, Yunrong Guo, Ofir Nabati, Gal Chechik, and Xue Bin Peng. MaskedMimic: Unified physics-based character control through masked motion inpainting. *ACM Transactions on Graphics (TOG)*, 2024. 2, 3, 4, 5, 6, 7, 8, 1

[49] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Amit H Bermano, and Daniel Cohen-Or. Human motion diffusion model. *arXiv*, 2022. 1, 2

[50] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. EDGE: Editable dance generation from music. In *CVPR*, 2023. 2

[51] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. HUMANISE: Language-conditioned human motion generation in 3D scenes. *NeurIPS*, 2022. 2, 3

[52] Yan Wu, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang. SAGA: Stochastic whole-body grasping with contact. In *ECCV*, 2022. 1, 2

[53] Zhen Wu, Jiaman Li, and C Karen Liu. Human-object interaction from human-level instructions. *arXiv*, 2024. 2

[54] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. OmniControl: Control any joint at any time for human motion generation. *arXiv*, 2023. 2, 4, 5, 6, 1

[55] 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

[56] Sirui Xu, Ziyin Wang, Yu-Xiong Wang, and Liang-Yan Gui. InterDreamer: Zero-shot text to 3D dynamic human-object interaction. *arXiv*, 2024. 2

[57] Sirui Xu, Dongting Li, Yucheng Zhang, Xiyan Xu, Qi Long, Ziyin Wang, Yunzhi Lu, Shuchang Dong, Hezi Jiang, Akshat Gupta, Yu-Xiong Wang, and Liang-Yan Gui. InterAct: Advancing large-scale versatile 3D human-object interaction generation. In *CVPR*, 2025. 2, 3

[58] Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, and Liangyan Gui. InterMimic: Towards universal whole-body control for physics-based human-object interactions. In *CVPR*, 2025. 2

[59] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose++: Vision transformer for generic body pose estimation. *IEEE TPAMI*, 2023. 4, 5

[60] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *CVPR*, 2022. 2, 3

[61] Ye Yuan and Kris Kitani. DLow: Diversifying latent flows for diverse human motion prediction. In *ECCV*, 2020. 1

[62] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *CVPR*, 2023. 1, 2

[63] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. COUCH: Towards controllable human-chair interactions. In *ECCV*, 2022. 3

[64] Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3D bodies move. In *CVPR*, 2021. 1

[65] Kaifeng Zhao, Yan Zhang, Shaofei Wang, Thabo Beeler, and Siyu Tang. Synthesizing diverse human motions in 3D indoor scenes. In *ICCV*, 2023. 2

[66] Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu. Taming diffusion models for audio-driven co-speech gesture generation. In *CVPR*, 2023. 2