Perceiving and Acting in First-Person: A Dataset and Benchmark for Egocentric Human-Object-Human Interactions

Liang Xu[1,2,3] Chengqun Yang[1] Zili Lin[1,2,3] Fei Xu[1] Yifan Liu[1] Congsheng Xu[1]
Yiyi Zhang[4] Jie Qin[5] Xingdong Sheng[6] Yunhui Liu[6] Xin Jin[2,3] Yichao Yan[1*]
Wenjun Zeng[2,3] Xiaokang Yang[1]

[1]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[2]Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo, China
[3]Ningbo Key Laboratory of Spatial Intelligence and Digital Derivative, Ningbo, China
[4]MoE Key Lab of AI, School of Computer Science, Shanghai Jiao Tong University
[5]Nanjing University of Aeronautics and Astronautics [6]Lenovo

https://liangxuy.github.io/InterVLA/

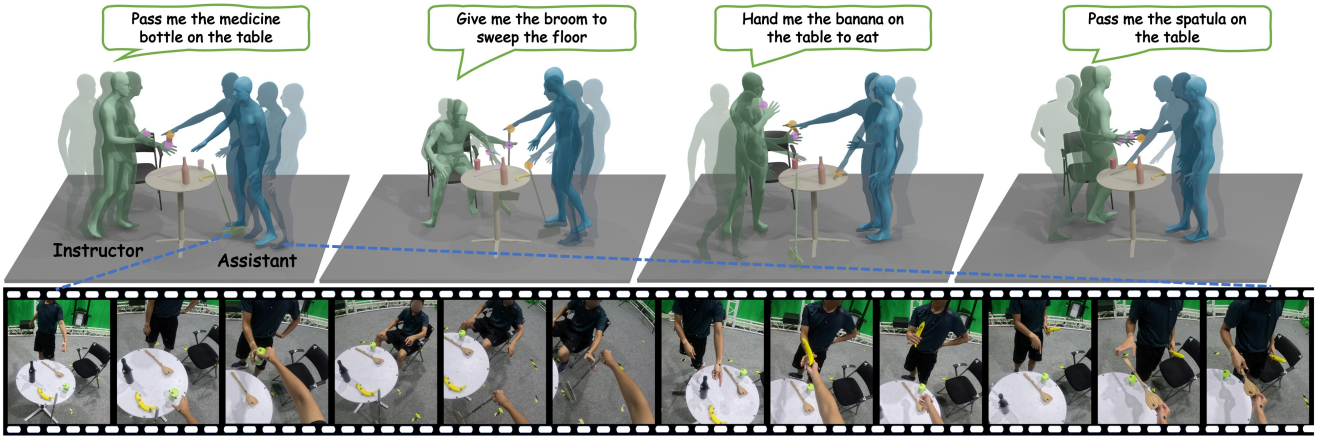


Figure 1. InterVLA features a large-scale human-object-human interaction dataset in a vision-language-action scheme, where an assistant provides services to an instructor based on egocentric perception and verbal commands. This comprehensive dataset comprises **3.9K** sequences, totaling **11.4** hours and **1.2M** frames of multimodal interaction data, including egocentric and exocentric RGB videos, language commands and high-precision human/object motions, promoting the development of general-purpose intelligent AI assistants.

Abstract

Learning action models from real-world human-centric interaction datasets is important towards building general-purpose intelligent assistants with efficiency. However, most existing datasets only offer specialist interaction category and ignore that AI assistants perceive and act based on first-person acquisition. We urge that both the generalist interaction knowledge and egocentric modality are indispensable. In this paper, we embed the manual-assisted task into a vision-language-action framework, where the assistant provides services to the instructor following egocentric vision and commands. With our hybrid RGB-MoCap system, pairs of assistants and instructors engage with multiple objects and the scene following GPT-generated scripts. Under this setting, we accomplish InterVLA, the first large-scale human-object-human interaction dataset with 11.4 hours and 1.2M frames of multimodal data, spanning 2 egocentric and 5 exocentric videos, accurate human/object motions and verbal commands. Furthermore, we establish novel benchmarks on egocentric human motion estimation, interaction synthesis, and interaction prediction with comprehensive analysis. We believe that our InterVLA testbed and the benchmarks will foster future works on building AI agents in the physical world.

1. Introduction

Learning generalist interaction knowledge is indispensable towards general-purpose intelligent agents to assist humans in the physical world. To avoid expensive robot data col-

*Corresponding authors

lection, learning from human-centric interactive datasets is more efficient [8, 77, 80, 97]. Existing datasets on human-human interactions contribute to human-robot interactions [14, 56, 81, 93], teleoperation [44, 45, 100]; human-object interactions promote human-to-robot handover [17, 18, 115, 116], human-robot collaboration [20, 98, 134] and human-scene interactions advance navigation [118, 136, 139].

Despite the rapid development of magnitude and richness in human-centric datasets and benchmarks, they still face some limitations in building intelligent assistants. Imagining the most basic capabilities for home robots, perceiving and comprehending the instructor's commands, navigating smoothly, and manipulating objects are required. However, most datasets only offer specialist interaction category [37, 66, 76, 86, 116, 122, 130, 134] rather than a generic scenario composed of diverse human-human, object and scene interactions. Besides, existing datasets [37, 66, 76, 116, 122, 130, 137] ignore the fact that AI assistants always perceive and then react based on their first-person acquisition [112, 134]. The absence of egocentric perspective could hinder the physical deployment of AI assistants.

To address these limitations and foster the development of general human-centric interactions and versatile AI assistants, a comprehensive dataset encompassing **diverse interaction patterns** and **stable egocentric perception** is pivotal. In this paper, we focus on the common daily scenarios of manual-assisted tasks with the majority being human-object-human interactions where an assistant providing services to an instructor following the egocentric vision and verbal commands, such as *"Pass me the cup on the table"*, where the human-human, object, scene interactions are naturally integrated. To simulate real-world robotic assistance scenarios, we randomly arrange various pieces of furniture and additional operable objects to establish the scene. The instructor gives verbal commands accompanied by body gestures while the assistant comprehends the intention and then responds accordingly.

Inspired by the vision-language-action (VLA) paradigm emerged for instruction-following robots, we formulate our data collection setup within the VLA framework and introduce InterVLA, the first large-scale egocentric human-object-human interaction dataset with various interaction categories as depicted in Fig. 1. For the *vision* modality, we capture two egocentric videos from the instructor's perspective and five exocentric videos covering the full scene. The *language* component consists of 100 meticulously crafted scripts featuring various scene arrangements, versatile interaction types, multi-object interactions and navigation tasks. To acquire *action* data, we attach reflective markers to the human and object surfaces, enabling high-precision motion tracking while preserving RGB data fidelity. We recruit **47** participants to form **27** unique instructor-assistant pairs engaging with **50** household objects, yielding **3.9K** sequences of **11.4** hours and **1.2M** frames of interaction data in total. All the captured data are well-calibrated and temporally synchronized. A comparison with existing human-centric interactive datasets is summarized in Tab. 1.

Dataset	Modality						Scale		
	HHI	HOI	HSI	Multi-Obj	Ego	Exo	#Seqs	#Objs	#Hours
You2Me [86]	✓	✗	✗	✗	✓	✗	42	-	1.4
ExPI [37]	✓	✗	✗	✗	✗	✓	115	-	0.3
Hi4D [130]	✓	✗	✗	✗	✗	✓	100	-	-
InterHuman [66]	✓	✗	✗	✗	✗	✓	6.0K	-	6.6
Inter-X [122]	✓	✗	✗	✗	✗	✗	11.4K	-	18.8
HIMO [76]	✗	✓	✗	✓	✗	✗	3.4K	53	9.4
HOH [116]	✗	✓	✗	✗	✗	✓	2.7K	136	-
CORE4D [134]	✓	✓	✗	✗	✓	✓	1.0K	37	-
HOI-M^3 [137]	✓	✓	✓	✓	✗	✓	199	90	20
InterVLA	✓	✓	✓	✓	✓	✓	3.9K	50	11.2

Table 1. **Dataset comparison.** We compare InterVLA with existing human-centric interactive datasets. **Modality** measures the human-human, human-object, human-scene interactions, multi-object interactions, egocentric and exocentric views. **Scale** measures the number of sequences, objects and hours.

With our proposed InterVLA dataset, we introduce novel tasks and benchmarks on how AI assistants better perceive the surroundings and then generate appropriate responses. We formulate four downstream tasks of 1) Egocentric human motion estimation to extract the global body motion of the instructor based on the first-person perspective of the assistant, 2) Interaction synthesis to generate plausible human-object-human sequences given the textual descriptions and the initial states of the human and objects, 3) Motion-based interaction prediction to anticipate the future human/object motions conditioned on the previous motion frames and 4) Vision-language based interaction prediction to forecast the future human motions from the historical first-person videos and the verbal instruction. We establish comprehensive benchmarks for these tasks, providing baseline models, quantitative and qualitative evaluations, and in-depth analysis. The results highlight the challenges posed by rapid camera movement, limited visibility, occlusions, and multi-object interactions within InterVLA. Additionally, we emphasize the dataset's potential applications in tasks such as sparse-view 4D scene reconstruction, hand-object interaction, and motion reconstruction from sparse signals. Our contributions can be summarized as:

- We collect the first large-scale human-object-human interaction dataset called InterVLA with diverse generalist interaction categories and egocentric perspectives.
- Our proposed benchmark with thorough analysis on egocentric human motion estimation, interaction synthesis and interaction prediction will stimulate future works on building intelligent AI assistants. We will release all the datasets, code and models for further exploration.

2. Related Work

Egocentric Vision. Increasing attention is attached to egocentric vision with datasets [6, 22, 23, 32, 34, 35, 41, 62, 64,

(a) Human Setup (b) Object-Scene Setup (c) Hybrid RGB-MoCap System

Figure 2. **InterVLA capturing system.** (a). We attach 41 reflective markers to the subject's body with strong medical glue to track the human body motion. The *assistant* wears two GoPro cameras to capture the egocentric data. (b). Reflective markers are positioned on the surface of real objects to track the precise object trajectories. We also collect the precise 3D object scans with a KSCAN Magic Scanner. (c). Our hybrid RGB-MoCap system with two egocentric RGB cameras, five exocentric RGB cameras, and an OptiTrack MoCap system.

85, 112], spurred by applications like robotics, AR and VR. The unique dynamics and perspective of egocentric vision present new challenges and opportunities for various tasks, including wearer pose estimation [3, 53, 62, 75, 82, 110, 111], activity recognition [35, 57, 65, 141], human-object interaction (HOI) [15, 21, 61, 77, 84], robotic active perception [4, 58, 107, 126], and interactive assistants [112]. InterVLA captures the egocentric data via two GoPro cameras showing in Fig. 2, together with accurate humans and objects motions obtained by the OptiTrack MoCap system. Compared to most video datasets such as Ego-Exo4D [35], InterVLA provides ground-truth 4D human and object motions captured by an optical MoCap system.

Human-Human Interactions. Besides numerous single-human motion datasets [36, 50, 67, 69, 91, 96, 113, 121, 143], several human-human datasets [33, 37, 66, 69, 86, 108, 122, 130, 133] have also been constructed for interaction synthesis [28, 66, 120, 122], human reaction generation [122, 123]. [14, 44, 56, 92, 93] also verifies that the learned interaction knowledge can be applied to human-robot interactions. InterVLA is essentially a human-human interaction dataset composed by an instructor and an assistant interacting with objects.

Human-Object Interactions. Recent efforts have expanded the boundaries of HOI datasets from hands-interactions [11, 39, 40, 51, 72, 73] and full-body interactions [29, 48, 105, 125] with single object to full-body interactions with multiple objects [76, 79]. In contrast to previous datasets focusing on individual HOI episode without context, InterVLA captures a series of coherent and consecutive HOI episodes for each GPT-generated script.

Human-Scene Interactions. Human-scene interactions incorporate comprehensive aspects ranging from navigation and collision avoidance to interaction with objects in the scene with various applications like embodied AI and VR. Real-world datasets [9, 38, 42, 43, 54, 83, 99, 138] face limitations such as static scenes [38, 43], single objects [9, 54, 138], or noisy 3D pose estimated from image [42, 83, 99]. Synthetic datasets [5, 10, 16, 55] resolve these issues yet suffer from appearance reality and physical plausibility. We build simple scenes with random furniture arrangements, offering comprehensive navigation and interaction data in real-world dynamic scenes.

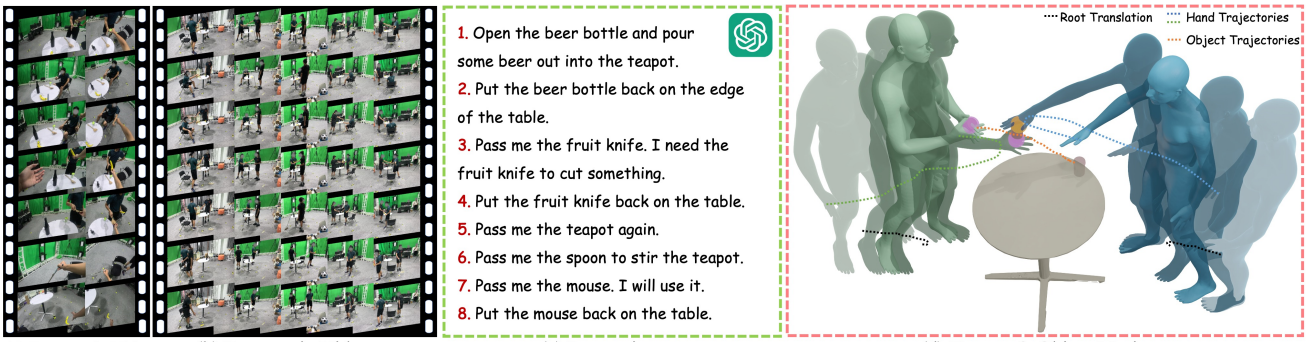
Human-Object-Human Interactions. HOH interactions are common in cooperative tasks and settings. Several datasets [17, 18, 27, 60, 116] focusing on handover are limited by fixed scene and body position settings, lack of egocentric RGB data, and absence of either natural human appearance or 3D human pose data. While CORE4D [134] provides egocentric HOH interaction data, it focuses exclusively on a single interaction type of two person collaboratively rearranging objects. InterVLA addresses those limitations to involve diverse HOH interaction types, flexible scene settings and multi-object interactions.

Vision-Language-Action Models. VLA models integrate vision and language as multi-modal inputs and generate actions for agents to accomplish tasks. The last few years have witnessed many great works [12, 13, 25, 47, 59, 78, 87, 140] propelling the advancement of VLA models with the support of large pre-trained models. HumanVLA [126] pioneeringly simulated humanoid VLA using simulator-style images as vision input. We anticipate InterVLA with real-world egocentric RGB data, language instructions, human response motion and object trajectories, could accelerate the practical humanoid VLA in real-world applications.

3. The InterVLA dataset

3.1. Overview

InterVLA is a large-scale human-object-human interaction dataset collected in a vision-language-action scheme, which



(a) Egocentric Videos (b) Exocentric Videos (c) Instructions (d) Human & Object Motions

Figure 3. **Components of the InterVLA dataset.** For the **vision** modality, we capture (a) two egocentric and (b) five exocentric RGB videos; For the **language** modality, we comprehensively supply the (c) GPT-generated commands; For the **action** modality, we provide the high-precision (d) human and object motions during the interactions.

features an assistant providing diverse services to the instructor in daily scenarios. We believe that our two-person and multi-object setting integrates several specialist human-centric interactions and will facilitate further research on robot-centric interactions. Besides, our InterVLA dataset also emphasizes the utility of egocentric perception and the assistant's action based on it. As depicted in Fig. 1, each scene comprises randomly arranged furniture with two persons interacting with several daily objects. The assistant performs a series of consecutive actions following the language commands of the instructor, such as *"Give me the mug on the table"*. We provide multi-view exocentric (third-person) viewpoints and two egocentric (first-person) perspectives from the assistant. The human and object motions are obtained by the optical motion capture (MoCap) system. Next, we will describe the vision-language-action data collection scheme and pipeline in Sec. 3.2 and dataset post-processing, components, and statistics in Sec. 3.3.

3.2. Vision-Language-Action Capturing

The concept of vision-language-action (VLA) emerges with the rise of instruction-following robotics. Inspired by it, we embed the manual-assisted task into the VLA framework where the assistant performs diverse services to the instructor such as picking up, retrieval, handover, and rearrangements of multiple objects. The hybrid RGB-MoCap capturing system of InterVLA is elaborately illustrated in Fig. 2.

Capturing Pipeline. We recruit 46 participants to form 27 unique instructor-assistant pairs for data collection. For the object and scene setup, we primarily select 50 common household objects of various sizes, including small objects such as fruit, mug, knife, and large objects such as suitcases, floor hangers, and besom. Note that we adopt real objects rather than 3D printed objects as [76, 105] for the fidelity of the RGB modality. The details of the object list are provided in the supplementary material. To align with real-world intelligent robotic assistance and enrich the interaction categories, we meticulously develop the following data-capturing pipeline. We first randomly arrange some furniture of tables or chairs in the MoCap venue with several operable objects positioned in the scene. During collection, the instructor first communicates with the assistant using verbal commands along with complementary body gestures, while the assistant should interpret the instructor's intention and react appropriately. For each recording, a sequence of atomic interactions is performed to preserve long-duration interactions and maintain continuity.

Vision. Two egocentric GoPro cameras are mounted tightly on the forehead and chest of the *assistant*, respectively, to capture first-person RGB videos with a high resolution of 5312×2988 at 30 fps. The camera intrinsics are obtained following [1]. The camera positions and orientations are carefully adjusted based on the height of the participants to maximize the valid shooting area of the scene, the other person, and the interactions. Besides, we also integrate five well-calibrated RGB cameras to compensate for the multi-view exocentric viewpoints with a resolution of 1920×1080 at 30 fps. The egocentric and exocentric cameras are all temporally synchronized by millisecond-level timestamps. To protect the privacy of participants, we mask the faces of all the exocentric and egocentric RGB videos with [2].

Language. Commands of the instructor serve as the starting point, trigger, and bridge for the following interactions. Given the household objects and furniture library, we employ large language models, *i.e.*, ChatGPT [88] to select the scenes and objects setup, determine their placement, and then generate a script of instructor-assistant interaction sequences within the scene based on the object affordance. With the majority of interactions focusing on human-object-human interactions such as handover and collaborative rearrangement, we also include some pure human-human interactions such as supporting or massaging. Furthermore, we encourage multi-object interactions with object-object interactions such as *"Slicing the apple with a knife"* and simultaneous manipulation such as *"Tidying up the objects*

on the table". Additionally, moving around and navigating within the scene are also supported. Ultimately, we produce 100 scripts, each involving an average of 2-3 furniture, 5 household objects, and 8 consecutive atomic commands. Each script is manually reviewed to ensure validity.

Action. For robotic arms, *action* can be defined as the rotations and translations of the robot joints. Similarly, motion is a compact representation for modeling human actions and object movements. To acquire high-quality human and object motions, we establish a MoCap system of 8.5×5.4 meters with 20 infrared cameras. For the human motions, we discard the tight MoCap suits but instead directly attach the reflective markers on the skin or clothes surface by strong medical glue as [29, 49] to preserve the fidelity of the RGB modality. The relative displacement between skin and clothing is eliminated by glue to ensure robust and precise MoCap results. The objects are treated as rigid bodies with at least four reflective markers placed on the surfaces for optical tracking. In our setting, we adopt the 12.5mm diameter reflective spheres for both humans and objects to achieve the best tracking results. All the assets are well-created and calibrated before recording. We also equip the MoCap system with timecodes for the temporal alignment with the ego-exo RGB videos.

3.3. Dataset Components

Human Parametric Model. SMPL parametric model [74] is widely adopted in human-centric interaction datasets, which formulates the human mesh as the body pose $\theta \in \mathbb{R}^{23\times3}$, global orientation $q_i \in \mathbb{R}^3$, root translation $\gamma_i \in \mathbb{R}^3$ and the body shape parameters $\beta \in \mathbb{R}^{10}$, which are determined based on the height, weight and gender of the participant as [76, 94, 122]. We fit the BVH-format human skeleton captured from the MoCap data to the SMPL parameters with the following optimization objective as:

$$\mathcal{L} = \lambda_j \mathcal{L}_j + \lambda_s \mathcal{L}_s + \lambda_{reg} \mathcal{L}_{reg}, \tag{1}$$

where $\mathcal{L}_j$ measures the difference between the raw MoCap joint position and the optimized result, $\mathcal{L}_s$ smooths the inter-frame motion transitions and mitigates pose jittering, $\mathcal{L}_{reg}$ regularizes the optimized poses from deviating and $\lambda_j = 1$, $\lambda_s = 0.1$, $\lambda_{reg} = 0.01$ are loss weights. We give further details of each optimization term in supplementary.

Object Meshes and Tracking. To render accurate human-object interactions, we scan all 50 objects and obtain the precise object surface geometries with a KSCAN Magic Scanner. Each object is then attached by more than 3 reflective markers and tracked by the optical MoCap system. We further scan the objects together with the attached markers and align the new scans with the previous results to eliminate the offsets between the two centroids. The object motion can be represented as the translation $t^o \in \mathbb{R}^3$ and rotation $r^o \in \mathbb{R}^6$ of the 6D rotation representation [142].

Alignment and Segmentation. We standardize all the video data to a resolution of 1920×1080 at 30 fps and downsample the MoCap data to the same fps for consistency. The exocentric videos, egocentric videos and motion data are well synchronized with millisecond-level timestamps. To facilitate training for downstream tasks, we render the interaction results and then manually split the long-duration script into short clips by the atomic commands while retaining the temporal continuity across the clips. The RGB videos are also segmented in the same way.

InterVLA Statistics. As aforementioned, we recruit 47 participants forming 27 unique instructor-assistant pairs for the data collection, spanning 519 valid long-duration interaction recordings with different GPT-generated scripts and scene arrangements. We ensure that each object appears in at least five scenarios. After the alignment, segmentation and thorough manual refinement of the motion data, we obtain 3,906 interaction sequences with 11.4 hours and 1.2M frames of four-tuple data composed by the egocentric videos, exocentric videos, instructions and human/object motions. We showcase an example of the dataset in Fig. 3 for a better illustration and more video examples will be provided in the supplementary materials.

4. Tasks, Benchmarks and Experiments

We define new tasks and benchmarks of egocentric human motion estimation, interaction synthesis and interaction prediction as presented in Fig. 4 and potential downstream tasks based on InterVLA, oriented towards general-purpose AI agents. We also provide the evaluation results for the proposed benchmarks with detailed analysis.

4.1. Preliminary Formulation

Egocentric videos exhibit the active perspective that shows the superiority of capturing the details of the close-by interactions and the first-person intentions. However, they also suffer from the constrained camera field of view without global perception, rapid viewpoint changes of objects exiting and re-entering the field of the frame.

We formulate our multimodal data as $[\boldsymbol{V}, \boldsymbol{T}, \boldsymbol{H}, \boldsymbol{O}, \boldsymbol{l}]$ for each sequence, where $\boldsymbol{V} = \{\boldsymbol{v}_h, \boldsymbol{v}_c\}$ are the **h**ead-mounted and **c**hest-mounted egocentric videos respectively, $\boldsymbol{T} = \{\boldsymbol{t}_i\}_{i=0}^{N_t}$ are the exocentric videos and $N_t = 5$ indicates the video number. The human motions are denoted as $\boldsymbol{H} = \{\boldsymbol{h}_I, \boldsymbol{h}_A\}$ for the **I**nstructor and **A**ssistant respectively, where $\boldsymbol{h}_I/\boldsymbol{h}_A$ can be detailed as $\{\theta \in \mathbb{R}^{23\times6}, q \in \mathbb{R}^6, \gamma \in \mathbb{R}^3, \beta \in \mathbb{R}^{10}\}$ for the body pose, global orientation, root translation and the body shape parameters, respectively. The object motions can be represented as $\boldsymbol{O} = \{\boldsymbol{o}_i\}_{i=0}^{N_o}$, where $\boldsymbol{o}_i = \{r_i^o \in \mathbb{R}^6, t_i^o \in \mathbb{R}^3\}$ for the rotation and translation of the object, respectively. N_o means the number of objects. The geometries of all involved objects are precisely scanned denoted as $\boldsymbol{G} = \{\boldsymbol{g}_i\}_{i=0}^{N_o}$. Here

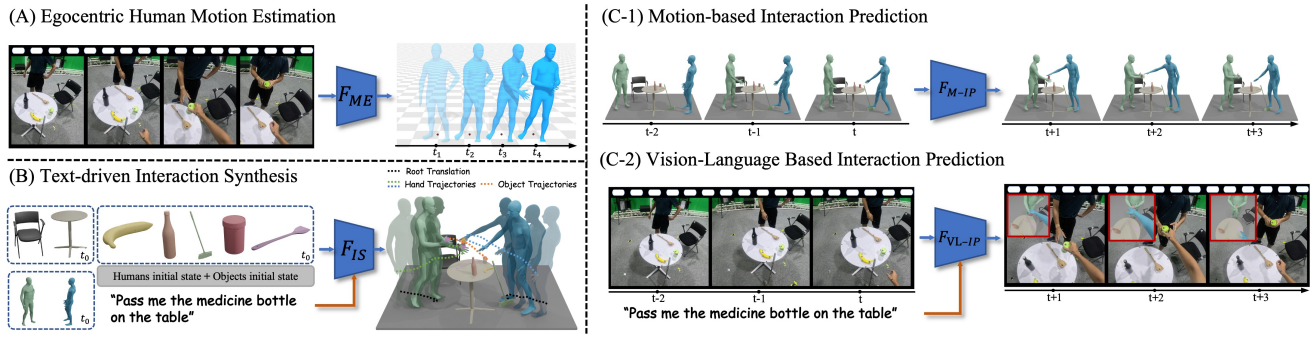


Figure 4. **Task Formulation of InterVLA.** We establish multiple downstream tasks on egocentric human motion estimation, text-driven interaction synthesis, motion-based interaction prediction and vision-language based interaction prediction. All these benchmarks showcase the great challenges of our InterVLA dataset and will benefit practical, intelligent AI assistants.

we adopt the 6D rotational representation [142] for humans and objects as in previous works [36, 90]. $\boldsymbol{l}$ refers to the language commands. We split the dataset into training, testing and validation sets with the ratio of 0.8, 0.15 and 0.05 for all downstream tasks. Note that we adopt **only the head-mounted camera** $\boldsymbol{v}_h$ for all the video-based experiments.

4.2. Egocentric Human Motion Estimation

Task Formulation. Egocentric perception and comprehension of the instructor's intention are fundamental as the first step of AI assistants. As depicted in Fig. 4 (A), we aim to reconstruct the world-grounded human motion sequences given the egocentric RGB videos as

$$F_{ME}(\boldsymbol{v}_h) \mapsto \hat{\boldsymbol{h}}_I, \tag{2}$$

where $\hat{\boldsymbol{h}}_I$ denotes the estimated instructor motion. Note that it is quite challenging to maintain accurate body pose estimation and keep the consistent global coordinate system. The rapid movement of egocentric cameras, occlusion and limited visibility raise substantial challenges for this task.

Experiment Settings. We evaluate four state-of-the-art global human motion estimation methods, TRACE [104], GLAMR [132], TRAM [114], and WHAM [103] on all the head-mounted camera sequences of the InterVLA dataset for an intuitive understanding.

Evaluation Metrics. Following previous works [103, 114], We compute Mean Per Joint Position Error (MPJPE), Procrustes-aligned MPJPE (PA-MPJPE) and Per Vertex Error (PVE) to evaluate the 3D human pose and body shape estimation performance, and Acceleration error (Accel) for the inter-frame motion smoothness. More specifically, MPJPE calculates the average of the Euclidean distances between the predicted and ground truth joint positions. PA-MPJPE is the MPJPE calculated after Procrustes analysis that aligns the predicted poses to the ground truth through translation, rotation, and scaling. PVE measures the average distance between predicted and ground truth positions of the 6,890 vertices derived from the SMPL parametric model. Accel calculates the average difference in acceleration between the predicted and ground-truth coordinates.

Results and Analysis. We present the visualization results comparison in Fig. 5 with the raw egocentric frames, the ground truth motion captured by the MoCap system and the baseline results. Note that we manually add a horizontal offset to avoid the motion entanglement, yet we place the *anchor points* on the floor to refer the global translation of the motion. The results show that even the best-performing algorithms obtain unsatisfactory results. For the three samples, the assistant turns head to locate scene objects slightly (first-row) or rapidly (third-row), causing the instructor to move out of the camera's field of view. The red red dashed boxes show that GLAMR and WHAM fail to track the correct global orientation and keep the correct coordinate system. The green dashed boxes show that TRACE misses many frames. Besides, we also provide the quantitative results comparisons in Tab. 2, where WHAM [103] show remarkable superiority over other methods across all metrics except PA-MPJPE. However, there remains a significant gap between its performance and the ground truth. This discrepancy can be attributed to several challenges inherent in our dataset, such as occlusions, rapid camera motion movements, the incomplete capture of human bodies, and frequent occurrences of subjects entering and exiting the frame (re-entering).

4.3. Interaction Synthesis

Task Formulation. Following existing text-driven HOI synthesis methods [24, 76, 89, 137], we define the task as multiple humans and objects motion generation as shown in Fig. 4 (B). We formulate this task as

$$F_{IS}(\boldsymbol{H}^0, \boldsymbol{O}^0, \boldsymbol{G}, \boldsymbol{l}) \mapsto \{\hat{\boldsymbol{H}}, \hat{\boldsymbol{O}}\}, \tag{3}$$

where $\boldsymbol{H}^0$ and $\boldsymbol{O}^0$ represent the initial pose of the humans and objects respectively, and $\hat{\boldsymbol{H}}$ and $\hat{\boldsymbol{O}}$ are the generated human and object motions. Compared with previous HOI datasets such as BEHAVE [9], our InterVLA dataset con-

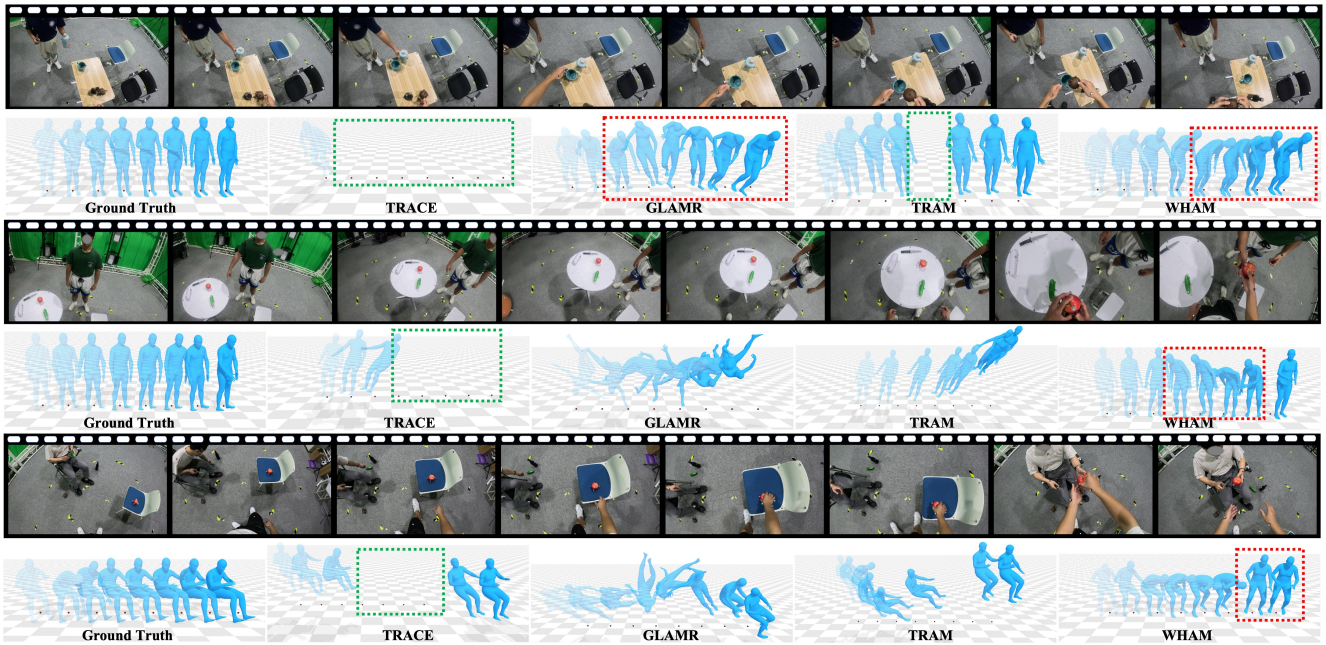


Figure 5. **Visualization result comparison** of the egocentric world-grounded human motion estimation results on the InterVLA dataset. For each sequence, we provide the original RGB frames along with the ground truth motion obtained by the MoCap system and four baseline models. Higher opacity indicates later frames of the sequence. Please zoom in for a more detailed view.

Method	PA-MPJPE↓	MPJPE↓	PVE↓	Accel↓
TRACE [104]	**91.2**	720.1	761.3	31.0
GLAMR [132]	134.8	589.9	596.7	24.9
TRAM [114]	102.7	684.1	718.7	27.5
WHAM [103]	103.2	**333.6**	**359.7**	**8.7**

Table 2. **Quantitative results** of egocentric global human motion estimation on the InterVLA dataset. **Bold** for the best results.

Methods	R Precision (Top 3) ↑	FID ↓	MM Dist ↓	Diversity →	MModality ↑
Real	$0.7592^{\pm 0.0026}$	$0.0203^{\pm 0.0018}$	$3.8718^{\pm 0.0064}$	$9.0164^{\pm 0.0831}$	–
MDM [106]	$0.4897^{\pm 0.0067}$	$2.8039^{\pm 0.0727}$	$5.4879^{\pm 0.0212}$	$7.7260^{\pm 0.0633}$	$1.9888^{\pm 0.0694}$
priorMDM [101]	$0.5250^{\pm 0.0068}$	$6.2766^{\pm 0.0777}$	$5.5129^{\pm 0.0188}$	$8.9414^{\pm 0.0888}$	$\mathbf{2.1227^{\pm 0.0993}}$
HIMO [76]	$\mathbf{0.5707^{\pm 0.0029}}$	$\mathbf{0.6805^{\pm 0.0136}}$	$\mathbf{4.9609^{\pm 0.0162}}$	$\mathbf{8.9849^{\pm 0.0554}}$	$1.1478^{\pm 0.0658}$

Table 3. **Quantitative results** of human-object-human interaction synthesis on the InterVLA dataset, where $\pm$ indicates 95% confidence interval and $\rightarrow$ means the closer the better. **Bold** highlights the best results.

tains multiple humans and objects and multi-object manipulations, which introduces greater challenges.

Experiment Settings. Text-driven human motion generation methods MDM [106], priorMDM [101] and HIMO [76] are re-implemented to support the condition input of object meshes and the initial states of two persons and multiple objects. Further details regarding these methods can be found in the supplementary materials.

Evaluation Metrics. We train the text feature extractor and human-object motion feature extractor first via contrastive learning as [36, 76]. The generation quality is evaluated by the following metrics: R Precision to evaluate the top-3 accuracy in retrieving the ground-truth description, Frechet Inception Distance (FID) [46] to measure the latent space divergence between authentic and synthetic samples, MultiModal distance (MM Dist) to determine the latent space distance between generated motions and input texts, Diversity to gauge the variance within the latent space, multimodality (MModality) to quantify the diversity of outputs generated from the same textual input.

Results and Analysis. The quantitative results presented in Tab. 3 demonstrate that HIMO [76] surpasses other methods across all metrics except for MModality, with a particularly impressive performance on FID. All these methods achieve a higher FID than the real interaction data, which shows that there remains ample opportunity for future endeavors to enhance the naturalness of the generated interaction results.

4.4. Interaction Prediction

Task Formulation. We propose two types of interaction prediction tasks of motion-based and vision-language based as demonstrated in Fig. 4 (C1) and (C2). For the motion-based interaction prediction, the model predicts the subsequent HOI sequences for the following frames given the adjacent past few frames of HOI sequences as

$$F_{M\text{-}IP}(\boldsymbol{H}^{t_I:t-1}, \boldsymbol{O}^{t_I:t-1}, \boldsymbol{G}) \mapsto \{\hat{\boldsymbol{H}}^{t:t_E}, \hat{\boldsymbol{O}}^{t:t_E}\}, \quad (4)$$

Methods	Human	Object		Contact	
	J_e(mm, ↓)	T_e(mm, ↓)	R_e(°, ↓)	C_{acc}(%, ↑)	P_r(%, ↓)
MDM [106]	175.3 (± 0.8)	140.2 (±0.7)	11.0 (±0.2)	85.5 (±0.3)	0.4 (±0.0)
InterDiff [124]	175.3 (± 0.8)	138.7 (±0.6)	10.8 (±0.1)	86.0 (±0.2)	0.4 (±0.0)
CAHMP [19]	**172.5 (± 0.4)**	**115.6 (±0.5)**	**9.5 (±0.1)**	-	-

Table 4. **Quantitative results** of motion-based interaction prediction on the InterVLA dataset.

where t_I (t_E) denotes the initial (ending) frame of the sequence, $\hat{\boldsymbol{H}}^{t:t_E}$ and $\hat{\boldsymbol{O}}^{t:t_E}$ are the predicted future motions. In the experiments, we set $t_I = 0$, $t = 15$ and $t_E = 30$ to predict the poses in the subsequent 15 frames given the previous 15 frames. While for the vision-language based interaction prediction, the model manages to anticipate the future human motions from the historical egocentric videos and the verbal instruction as

$$F_{VL\text{-}IP}(\boldsymbol{v}_h{}^{t_I:t-1}, \boldsymbol{h}_A^{t_I:t-1}, \boldsymbol{l}) \mapsto \{\hat{\boldsymbol{h}_A}^{t:t_E}\}, \tag{5}$$

where $\hat{\boldsymbol{h}_A}^{t:t_E}$ means the predicted motion of the assistant.

Experiment Settings. We evaluate three state-of-the-art methods CAHMP [19], MDM [106] and InterDiff [124] for motion-based interaction prediction. For vision-language guided interaction prediction, we re-implement the existing hand trajectory prediction models FHOI [70], OCT [71], and USST [7] with integrated language embeddings to predict the motion of the assistant.

Evaluation Metrics. For motion-based interaction prediction, we follow [134] to apply the evaluation metrics of human joint position error (J_e) to measure the Mean Per Joint Position Error (MPJPE) for two individuals; object translation error (T_e) representing the average $L2$ differences between predicted and real object translations; object rotation error (R_e), indicating the average geodesic differences between predicted and actual object rotations; human-object contact accuracy (C_{acc}) to assess the average error rate in contact detection with a 5cm threshold to detect hand contacts; and penetration rate (P_r) calculating the percentage of object vertices penetrating human meshes. For vision-language guided interaction prediction, we adopt the average displacement error as the average $L2$ distance between the predicted and ground truth trajectories, and the final displacement error as the $L2$ distance between the two final predicted and ground truth locations following FHOI [70].

Results and Analysis. We provide the quantitative results of three state-of-the-art models for motion-based interaction prediction in Tab. 4. From the results, we can derive that CAHMP [19] achieves the best performance over the other baseline methods for both the human joint position error and the object translation and rotation error, thanks to the semantic-graph model to learn the relationship between human and context objects. From the quantitative comparisons of the vision-language based interaction prediction in Tab. 5, we can derive that USST [7] achieves the best performance for the two metrics due to the proposed uncertainty-aware state space Transformer. However, we find that existing state-of-the-art methods have not achieved satisfactory performance on these two tasks. Significant errors remain in the predicted human motion, object translation & rotation and human-object contact for these two tasks, which indicates that our dataset presents significant challenges for subsequent optimization.

Method	Avg. Disp. Error ↓	Final Disp. Error ↓
FHOI [70]	0.29	0.38
OCT [71]	0.28	0.36
USST [7]	**0.24**	**0.32**

Table 5. **Quantitative results** of vision-language based interaction prediction on the InterVLA dataset.

4.5. Potential Downstream Tasks

Sparse-view 4D Scene Reconstruction. Multi-view exocentric videos are synergistic with the egocentric viewpoint for global awareness of the scene. Existing works [63, 109, 119, 131, 135] are dedicated to reconstructing static scenes from sparse-view images with 3D representations like meshes, neural radiance fields, and 3D Gaussian splatting while the majority of existing 4D reconstruction efforts [68, 95, 102, 117] are confined to dense perspective input. Our dataset comprising five exocentric-view videos, is anticipated to assist in downstream tasks of 4D scene reconstruction through additional supervisory signals, such as poses and meshes of people and objects.

Hand-object Interaction Reconstruction. Similar to [30, 31, 127, 128] that jointly estimate the poses of both hands and the interacting objects, our InterVLA dataset with egocentric viewpoints and multi-object manipulation can also empower this task with substantial challenges.

Motion Reconstruction from Sparse Signals. This task aims to reconstruct one's own whole-body motion of the assistant given the sparse signals of the egocentric captured low body or arms as in previous works [26, 52, 129].

5. Conclusion

In this paper, we introduce a large-scale egocentric human-object-human interaction dataset called InterVLA. By embedding the manual-assisted tasks into a vision-language-action scheme, we formulate *vision* as the egocentric perspective, *language* as the instructor's verbal commands and *action* as the human and object motions, and demonstrate the indispensability of both generalist interaction knowledge and egocentric perception for building physical-world AI assistants. Through our extensive dataset and novel benchmarks for egocentric motion estimation, interaction synthesis and interaction prediction, we provide valuable tools that will drive further research and development in the field of real-world AI-assisted applications.

Acknowledgements

This work was supported in part by NSFC (62201342), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), Grants of NSFC 62302246, ZJNSFC LQ23F010008, Ningbo 2023Z237 & 2024Z284 & 2024Z289 & 2023CX050011 & 2025Z038, and supported by High Performance Computing Center at Eastern Instituteof Technology and Ningbo Institute of Digital Twin. Authors would like to appreciate the Student Innovation Center of SJTU for providing GPUs.

References

[1] Easymocap - make human motion capture easier. https://github.com/zju3dv/EasyMocap. 4

[2] Opencv: opencv. https://github.com/opencv/opencv/blob/master/data/haarcascades/haarcascade_frontalface_default.xml. 4

[3] Hiroyasu Akada, Jian Wang, Vladislav Golyanik, and Christian Theobalt. 3d human pose perception from egocentric stereo videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 767–776, 2024. 3

[4] Boshi An, Yiran Geng, Kai Chen, Xiaoqi Li, Qi Dou, and Hao Dong. Rgbmanip: Monocular image-based robotic manipulation through active object pose estimation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7748–7755. IEEE, 2024. 3

[5] Joao Pedro Araújo, Jiaman Li, Karthik Vetrivel, Rishi Agarwal, Jiajun Wu, Deepak Gopinath, Alexander William Clegg, and Karen Liu. Circle: Capture in rich contextual environments. In *CVPR*, pages 21211–21221, 2023. 3

[6] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Fan Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, et al. Introducing hot3d: An egocentric dataset for 3d hand and object tracking. *arXiv preprint arXiv:2406.09598*, 2024. 2

[7] Wentao Bao, Lele Chen, Libing Zeng, Zhong Li, Yi Xu, Junsong Yuan, and Yu Kong. Uncertainty-aware state space transformer for egocentric 3d hand trajectory forecasting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13702–13711, 2023. 8

[8] Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation, 2024. 2

[9] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15935–15946, 2022. 3, 6

[10] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *CVPR*, pages 8726–8737, 2023. 3

[11] Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 361–378. Springer, 2020. 3

[12] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022. 3

[13] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. 3

[14] Judith Bütepage, Ali Ghadirzadeh, Özge Öztimur Karadağ, Mårten Björkman, and Danica Kragic. Imitating by generating: Deep generative models for imitation of interactive tasks. *Frontiers in Robotics and AI*, 7:47, 2020. 2, 3

[15] Minjie Cai, Kris M Kitani, and Yoichi Sato. Understanding hand-object manipulation with grasp types and object attributes. In *Robotics: Science and Systems*, 2016. 3

[16] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 387–404. Springer, 2020. 3

[17] Alessandro Carfì, Francesco Foglino, Barbara Bruno, and Fulvio Mastrogiovanni. A multi-sensor dataset of human-human handover. *Data in brief*, 22:109–117, 2019. 2, 3

[18] Francesca Cini, V Ortenzi, P Corke, and MJSR Controzzi. On the choice of grasp type and location when handing over an object. *Science Robotics*, 4(27):eaau9757, 2019. 2, 3

[19] Enric Corona, Albert Pumarola, Guillem Alenyà, and Francesc Moreno-Noguer. Context-aware human motion prediction, 2020. 8

[20] Ana Cunha, Flora Ferreira, Emanuel Sousa, Luis Louro, Paulo Vicente, Sergio Monteiro, Wolfram Erlhagen, and Estela Bicho. Towards collaborative robots as intelligent co-workers in human-robot joint tasks: what to do and who does it? In *ISR 2020; 52th International Symposium on Robotics*, pages 1–8. VDE, 2020. 2

[21] Dima Damen, Teesid Leelasawassuk, and Walterio Mayol-Cuevas. You-do, i-learn: Egocentric unsupervised discovery of objects and their modes of interaction towards video-based guidance. *Computer Vision and Image Understanding*, 149:98–112, 2016. 3

[22] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018. 2

[23] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2022. 2

[24] Christian Diller and Angela Dai. Cg-hoi: Contact-guided 3d human-object interaction generation. *arXiv preprint arXiv:2311.16097*, 2023. 6

[25] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 3

[26] Yuming Du, Robin Kips, Albert Pumarola, Sebastian Starke, Ali Thabet, and Artsiom Sanakoyeu. Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In *CVPR*, 2023. 8

[27] Tair Faibish, Alap Kshirsagar, Guy Hoffman, and Yael Edan. Human preferences for robot eye gaze in human-to-robot handovers. *International Journal of Social Robotics*, 14(4):995–1012, 2022. 3

[28] Ke Fan, Shunlin Lu, Minyue Dai, Runyi Yu, Lixing Xiao, Zhiyang Dou, Junting Dong, Lizhuang Ma, and Jingbo Wang. Go to zero: Towards zero-shot motion generation with million-scale data. *arXiv preprint arXiv:2507.07095*, 2025. 3

[29] 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 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12943–12954, 2023. 3, 5

[30] Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Xu Chen, Muhammed Kocabas, Michael J Black, and Otmar Hilliges. Hold: Category-agnostic 3d reconstruction of interacting hands and objects from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 494–504, 2024. 8

[31] Zicong Fan, Takehiko Ohkawa, Linlin Yang, Nie Lin, Zhishan Zhou, Shihao Zhou, Jiajun Liang, Zhong Gao, Xuanyang Zhang, Xue Zhang, et al. Benchmarks and challenges in pose estimation for egocentric hand interactions with objects. In *European Conference on Computer Vision*, pages 428–448. Springer, 2025. 8

[32] Alircza Fathi, Jessica K Hodgins, and James M Rehg. Social interactions: A first-person perspective. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1226–1233. IEEE, 2012. 2

[33] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *CVPR*, pages 7214–7223, 2020. 3

[34] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 2

[35] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 2, 3

[36] 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*, pages 5152–5161, 2022. 3, 6, 7

[37] Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Multi-person extreme motion prediction. In *CVPR*, pages 13053–13064, 2022. 2, 3

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

[39] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3196–3206, 2020. 3

[40] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11090–11100, 2022. 3

[41] Shengyu Hao, Wenhao Chai, Zhonghan Zhao, Meiqi Sun, Wendi Hu, Jieyang Zhou, Yixian Zhao, Qi Li, Yizhou Wang, Xi Li, et al. Ego3dt: Tracking every 3d object in ego-centric videos. *arXiv preprint arXiv:2410.08530*, 2024. 2

[42] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *ICCV*, pages 2282–2292, 2019. 3

[43] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *ICCV*, pages 11374–11384, 2021. 3

[44] Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. *arXiv preprint arXiv:2406.08858*, 2024. 2, 3

[45] Tairan He, Zhengyi Luo, Wenli Xiao, Chong Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Learning human-to-humanoid real-time whole-body teleoperation, 2024. 2

[46] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a

two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7

[47] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023. 3

[48] Yinghao Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. Intercap: Joint markerless 3d tracking of humans and objects in interaction. In *DAGM German Conference on Pattern Recognition*, pages 281–299. Springer, 2022. 3

[49] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 5

[50] Yanli Ji, Feixiang Xu, Yang Yang, Fumin Shen, Heng Tao Shen, and Wei-Shi Zheng. A large-scale rgb-d database for arbitrary-view human action recognition. In *ACMMM*, page 1510–1518, 2018. 3

[51] Juntao Jian, Xiuping Liu, Manyi Li, Ruizhen Hu, and Jian Liu. Affordpose: A large-scale dataset of hand-object interactions with affordance-driven hand pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14713–14724, 2023. 3

[52] Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *ECCV*, pages 443–460. Springer, 2022. 8

[53] Jiaxi Jiang, Paul Streli, Manuel Meier, and Christian Holz. Egoposer: Robust real-time ego-body pose estimation in large scenes. *arXiv preprint arXiv:2308.06493*, 2023. 3

[54] Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Zhiyuan Zhang, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9365–9376, 2023. 3

[55] 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

[56] Zhenyu Jiang, Yuqi Xie, Jinhan Li, Ye Yuan, Yifeng Zhu, and Yuke Zhu. Harmon: Whole-body motion generation of humanoid robots from language descriptions. *arXiv preprint arXiv:2410.12773*, 2024. 2, 3

[57] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5492–5501, 2019. 3

[58] Soheil Khatibi, Meisam Teimouri, and Mahdi Rezaei. Real-time active vision for a humanoid soccer robot using deep reinforcement learning. *arXiv preprint arXiv:2011.13851*, 2020. 3

[59] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. 3

[60] Alap Kshirsagar, Raphael Fortuna, Zhiming Xie, and Guy Hoffman. Dataset of bimanual human-to-human object handovers. *Data in Brief*, 48:109277, 2023. 3

[61] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10138–10148, 2021. 3

[62] Jiaman Li, Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17142–17151, 2023. 2, 3

[63] Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20775–20785, 2024. 8

[64] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European conference on computer vision (ECCV)*, pages 619–635, 2018. 2

[65] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6943–6953, 2021. 3

[66] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *arXiv preprint arXiv:2304.05684*, 2023. 2, 3

[67] 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. *arXiv preprint arXiv:2307.00818*, 2023. 3

[68] Youtian Lin, Zuozhuo Dai, Siyu Zhu, and Yao Yao. Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21136–21145, 2024. 8

[69] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *T-PAMI*, 42(10):2684–2701, 2019. 3

[70] Miao Liu, Siyu Tang, Yin Li, and James M Rehg. Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 704–721. Springer, 2020. 8

[71] Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots

prediction from egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3282–3292, 2022. 8

[72] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21022, 2022. 3

[73] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21740–21751, 2024. 3

[74] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6): 248:1–248:16, 2015. 5

[75] Zhengyi Luo, Ryo Hachiuma, Ye Yuan, and Kris Kitani. Dynamics-regulated kinematic policy for egocentric pose estimation. *Advances in Neural Information Processing Systems*, 34:25019–25032, 2021. 3

[76] Xintao Lv, Liang Xu, Yichao Yan, Xin Jin, Congsheng Xu, Shuwen Wu, Yifan Liu, Lincheng Li, Mengxiao Bi, Wenjun Zeng, et al. Himo: A new benchmark for full-body human interacting with multiple objects. In *European Conference on Computer Vision*, pages 300–318. Springer, 2025. 2, 3, 4, 5, 6, 7

[77] Junyi Ma, Jingyi Xu, Xieyuanli Chen, and Hesheng Wang. Diff-ip2d: Diffusion-based hand-object interaction prediction on egocentric videos. *arXiv preprint arXiv:2405.04370*, 2024. 2, 3

[78] Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. A survey on vision-language-action models for embodied ai. *arXiv preprint arXiv:2405.14093*, 2024. 3

[79] Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. Unifying representations and large-scale whole-body motion databases for studying human motion. *IEEE Transactions on Robotics*, 32(4):796–809, 2016. 3

[80] Priyanka Mandikal and Kristen Grauman. Dexvip: Learning dexterous grasping with human hand pose priors from video. In *Conference on Robot Learning*, pages 651–661. PMLR, 2022. 2

[81] Esteve Valls Mascaro, Daniel Sliwowski, and Dongheui Lee. Hoi4abot: Human-object interaction anticipation for human intention reading collaborative robots, 2024. 2

[82] Christen Millerdurai, Hiroyasu Akada, Jian Wang, Diogo Luvizon, Christian Theobalt, and Vladislav Golyanik. Eventego3d: 3d human motion capture from egocentric event streams. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1186–1195, 2024. 3

[83] Aron Monszpart, Paul Guerrero, Duygu Ceylan, Ersin Yumer, and Niloy J Mitra. imapper: interaction-guided scene mapping from monocular videos. *ACM Transactions On Graphics (TOG)*, 38(4):1–15, 2019. 3

[84] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8688–8697, 2019. 3

[85] Katsuyuki Nakamura, Serena Yeung, Alexandre Alahi, and Li Fei-Fei. Jointly learning energy expenditures and activities using egocentric multimodal signals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1868–1877, 2017. 3

[86] Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. You2me: Inferring body pose in egocentric video via first and second person interactions. In *CVPR*, pages 9890–9900, 2020. 2, 3

[87] Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023. 3

[88] OpenAI. GPT-3.5 turbo fine-tuning and api updates. `https://openai.com/index/gpt-3-5-turbo-fine-tuning-and-api-updates/`, 2023. 4

[89] 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. 6

[90] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *CVPR*, pages 10985–10995, 2021. 6

[91] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016. 3

[92] Vignesh Prasad, Dorothea Koert, Ruth Stock-Homburg, Jan Peters, and Georgia Chalvatzaki. Mild: multimodal interactive latent dynamics for learning human-robot interaction. In *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*, pages 472–479. IEEE, 2022. 3

[93] Vignesh Prasad, Alap Kshirsagar, Dorothea Koert Ruth Stock-Homburg, Jan Peters, and Georgia Chalvatzaki. Moveint: Mixture of variational experts for learning human-robot interactions from demonstrations. *IEEE Robotics and Automation Letters*, 2024. 2, 3

[94] Sergi Pujades, Betty Mohler, Anne Thaler, Joachim Tesch, Naureen Mahmood, Nikolas Hesse, Heinrich H Bülthoff, and Michael J Black. The virtual caliper: Rapid creation of metrically accurate avatars from 3d measurements. *IEEE transactions on visualization and computer graphics*, 25 (5):1887–1897, 2019. 5

[95] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 8

[96] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: bodies, action and behavior with english labels. In *CVPR*, pages 722–731, 2021. 3

[97] Yuzhe Qin, Yueh-Hua Wu, Shaowei Liu, Hanwen Jiang, Ruihan Yang, Yang Fu, and Xiaolong Wang. Dexmv: Imitation learning for dexterous manipulation from human videos. In *European Conference on Computer Vision*, pages 570–587. Springer, 2022. 2

[98] Ayumu Sasagawa, Kazuki Fujimoto, Sho Sakaino, and Toshiaki Tsuji. Imitation learning based on bilateral control for human–robot cooperation. *IEEE Robotics and Automation Letters*, 5(4):6169–6176, 2020. 2

[99] Manolis Savva, Angel X Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. Pigraphs: learning interaction snapshots from observations. *ACM Transactions On Graphics (TOG)*, 35(4):1–12, 2016. 3

[100] Mingyo Seo, Steve Han, Kyutae Sim, Seung Hyeon Bang, Carlos Gonzalez, Luis Sentis, and Yuke Zhu. Deep imitation learning for humanoid loco-manipulation through human teleoperation, 2023. 2

[101] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 7

[102] Meng-Li Shih, Jia-Bin Huang, Changil Kim, Rajvi Shah, Johannes Kopf, and Chen Gao. Modeling ambient scene dynamics for free-view synthesis. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 8

[103] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2070–2080, 2024. 6, 7

[104] Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J Black. Trace: 5d temporal regression of avatars with dynamic cameras in 3d environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8856–8866, 2023. 6, 7

[105] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *ECCV*, pages 581–600, 2020. 3, 4

[106] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 7, 8

[107] Dhruva Tirumala, Markus Wulfmeier, Ben Moran, Sandy Huang, Jan Humplik, Guy Lever, Tuomas Haarnoja, Leonard Hasenclever, Arunkumar Byravan, Nathan Batchelor, et al. Learning robot soccer from egocentric vision with deep reinforcement learning. *arXiv preprint arXiv:2405.02425*, 2024. 3

[108] NP Van der Aa, Xinghan Luo, Geert-Jan Giezeman, Robby T Tan, and Remco C Veltkamp. Umpm benchmark: A multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction. In *ICCV Workshops*, pages 1264–1269. IEEE, 2011. 3

[109] Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9065–9076, 2023. 8

[110] Jian Wang, Diogo Luvizon, Weipeng Xu, Lingjie Liu, Kripasindhu Sarkar, and Christian Theobalt. Scene-aware egocentric 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13031–13040, 2023. 3

[111] Jian Wang, Zhe Cao, Diogo Luvizon, Lingjie Liu, Kripasindhu Sarkar, Danhang Tang, Thabo Beeler, and Christian Theobalt. Egocentric whole-body motion capture with fisheyevit and diffusion-based motion refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 777–787, 2024. 3

[112] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, et al. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20270–20281, 2023. 2, 3

[113] Ye Wang, Sipeng Zheng, Bin Cao, Qianshan Wei, Qin Jin, and Zongqing Lu. Quo vadis, motion generation? from large language models to large motion models, 2024. 3

[114] 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*, pages 467–487. Springer, 2025. 6, 7

[115] Zifan Wang, Junyu Chen, Ziqing Chen, Pengwei Xie, Rui Chen, and Li Yi. Genh2r: Learning generalizable human-to-robot handover via scalable simulation demonstration and imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16362–16372, 2024. 2

[116] Noah Wiederhold, Ava Megyeri, DiMaggio Paris, Sean Banerjee, and Natasha Banerjee. Hoh: Markerless multimodal human-object-human handover dataset with large object count. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[117] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20310–20320, 2024. 8

[118] Zeqi Xiao, Tai Wang, Jingbo Wang, Jinkun Cao, Wenwei Zhang, Bo Dai, Dahua Lin, and Jiangmiao Pang. Unified human-scene interaction via prompted chain-of-contacts. *arXiv preprint arXiv:2309.07918*, 2023. 2

[119] Haolin Xiong, Sairisheek Muttukuru, Rishi Upadhyay, Pradyumna Chari, and Achuta Kadambi. Sparsegs: Real-time 360 {\deg} sparse view synthesis using gaussian splatting. *arXiv preprint arXiv:2312.00206*, 2023. 8

[120] Liang Xu, Ziyang Song, Dongliang Wang, Jing Su, Zhicheng Fang, Chenjing Ding, Weihao Gan, Yichao Yan, Xin Jin, Xiaokang Yang, et al. Actformer: A gan-based transformer towards general action-conditioned 3d human motion generation. In *ICCV*, pages 2228–2238, 2023. 3

[121] Liang Xu, Shaoyang Hua, Zili Lin, Yifan Liu, Feipeng Ma, Yichao Yan, Xin Jin, Xiaokang Yang, and Wenjun Zeng. Motionbank: A large-scale video motion benchmark with disentangled rule-based annotations, 2024. 3

[122] Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, et al. Inter-x: Towards versatile human-human interaction analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22260–22271, 2024. 2, 3, 5

[123] Liang Xu, Yizhou Zhou, Yichao Yan, Xin Jin, Wenhan Zhu, Fengyun Rao, Xiaokang Yang, and Wenjun Zeng. Regennet: Towards human action-reaction synthesis. In *CVPR*, pages 1759–1769, 2024. 3

[124] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. InterDiff: Generating 3d human-object interactions with physics-informed diffusion. In *ICCV*, 2023. 8

[125] Xiang Xu, Hanbyul Joo, Greg Mori, and Manolis Savva. D3d-hoi: Dynamic 3d human-object interactions from videos. *arXiv preprint arXiv:2108.08420*, 2021. 3

[126] Xinyu Xu, Yizheng Zhang, Yong-Lu Li, Lei Han, and Cewu Lu. Humanvla: Towards vision-language directed object rearrangement by physical humanoid. *arXiv preprint arXiv:2406.19972*, 2024. 3

[127] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3895–3905, 2022. 8

[128] Yufei Ye, Poorvi Hebbar, Abhinav Gupta, and Shubham Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19717–19728, 2023. 8

[129] Brent Yi, Vickie Ye, Maya Zheng, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. Estimating body and hand motion in an ego-sensed world, 2024. 8

[130] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *CVPR*, pages 17016–17027, 2023. 2, 3

[131] Hanyang Yu, Xiaoxiao Long, and Ping Tan. Lm-gaussian: Boost sparse-view 3d gaussian splatting with large model priors. *arXiv preprint arXiv:2409.03456*, 2024. 8

[132] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11038–11049, 2022. 6, 7

[133] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *CVPR*, pages 28–35. IEEE, 2012. 3

[134] Chengwen Zhang, Yun Liu, Ruofan Xing, Bingda Tang, and Li Yi. Core4d: A 4d human-object-human interaction dataset for collaborative object rearrangement. *arXiv preprint arXiv:2406.19353*, 2024. 2, 3, 8

[135] Chuanrui Zhang, Yingshuang Zou, Zhuoling Li, Minmin Yi, and Haoqian Wang. Transplat: Generalizable 3d gaussian splatting from sparse multi-view images with transformers. *arXiv preprint arXiv:2408.13770*, 2024. 8

[136] Hang Zhang, Wenxiao Zhang, Haoxuan Qu, and Jun Liu. Enhancing human-centered dynamic scene understanding via multiple llms collaborated reasoning. *Visual Intelligence*, 3(1):3, 2025. 2

[137] Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Hoi-mˆ 3: Capture multiple humans and objects interaction within contextual environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 516–526, 2024. 2, 6

[138] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: towards controllable human-chair interactions. In *ECCV*, pages 518–535, 2022. 3

[139] Xiaohan Zhang, Sebastian Starke, Vladimir Guzov, Zhensong Zhang, Eduardo Pérez Pellitero, and Gerard Pons-Moll. Scenic: Scene-aware semantic navigation with instruction-guided control. *arXiv preprint arXiv:2412.15664*, 2024. 2

[140] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: A 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*, 2024. 3

[141] Yipin Zhou and Tamara L Berg. Temporal perception and prediction in ego-centric video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4498–4506, 2015. 3

[142] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, pages 5745–5753. Computer Vision Foundation / IEEE, 2019. 5, 6

[143] Shihao Zou, Xinxin Zuo, Yiming Qian, Sen Wang, Chi Xu, Minglun Gong, and Li Cheng. 3d human shape reconstruction from a polarization image. In *ECCV*, pages 351–368, 2020. 3