FAVOR: Full-Body AR-Driven Virtual Object Rearrangement Guided by Instruction Text

Kailin Li[1*], Lixin Yang[1*], Zenan Lin[3], Jian Xu[2], Xinyu Zhan[1], Yifei Zhao[1], Pengxiang Zhu[1], Wenxiong Kang[3], Kejian Wu[2], Cewu Lu[1†]

[1]Shanghai Jiao Tong University
[2]XREAL
[3]South China University of Technology
{kailinli, siriusyang, kelvin34501, yifei_zhao, zhu_peng_xiang, lucewu}@sjtu.edu.cn,
{jianxu, kejian}@xreal.com, {auzenanlin, auwxkang}@mail.scut.edu.cn

Abstract

Rearrangement operations form the crux of interactions between humans and their environment. The ability to generate natural, fluid sequences of this operation is of essential value in AR/VR and CG. Bridging a gap in the field, our study introduces **FAVOR**: a novel dataset for Full-body AR-driven Virtual Object Rearrangement that uniquely employs motion capture systems and AR eyeglasses. Comprising 3k diverse motion rearrangement sequences and 7.17 million interaction data frames, this dataset breaks new ground in research data. We also present a pipeline **FAVORITE** for producing digital human rearrangement motion sequences guided by instructions. Experimental results, both qualitative and quantitative, suggest that this dataset and pipeline deliver high-quality motion sequences. Our dataset, code, and appendix are available at https://kailinli.github.io/FAVOR.

Introduction

Envision yourself immersed in augmented reality, arriving at a restaurant, your avatar utters, "Do me a **FAVOR**, make a burger for me". In response, the cook avatar precisely retrieves bread from the toaster, carefully placing it onto the plate. Subsequently, he gathers an array of ingredients and arranges them on top of the bread slice. Hence, for this application, it's crucial for the avatar to discern the relative positions of the on-scene objects, comprehend spatial preposition logic (such as 'on top of'), and accurately grasp and place the ingredients. Simply put, this involves *grounding* (text-scene alignment) and *rearrangement* (grasping-and-placement). The capability to generate such sequences of fluid and natural rearrangement actions is indispensable for the applications of digital human synthesis.

Over recent years, the field of full-body human motion generation has experienced rapid progression. For example, certain studies focused on body motion generation founded on language-based instruction (Tevet et al. 2022; Jiang et al. 2023), or on dynamics involving interaction with the scenes (Wang et al. 2022).

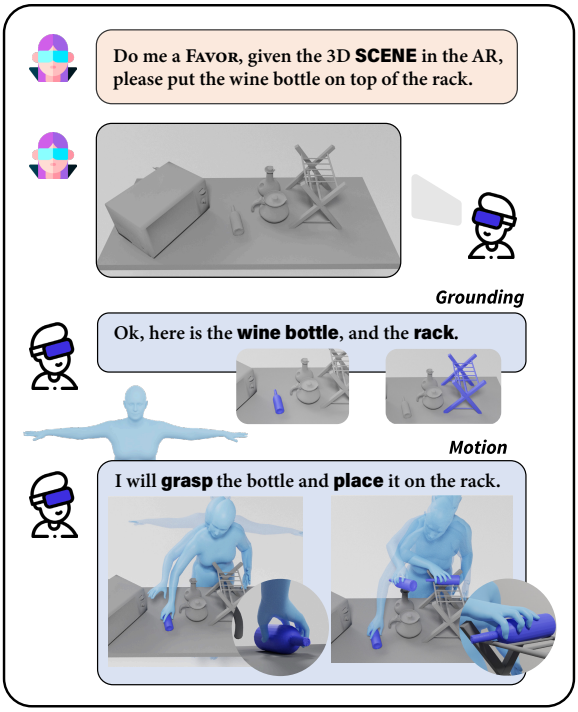

Figure 1: Illustration of the FAVOR data collection pipeline. The researcher directs the task through textual instructions and projects the scene onto the AR glasses. Subjects then rearrange the objects via interaction within the AR space.

However, the task of generating human-object interactions has been relatively less explored. Within this niche, the GRAB (Taheri et al. 2020) dataset serves as the benchmark for the generation of human-object motion. Nonetheless, its scope is limited to fundamental interactions (*e.g.* grasp-and-use) involving a human and a singular object, typically situated at the center of a desk. The lack of paired 'instruction-motion' sequence data hinders progress toward complex motion generation – specifically, tasks that require a nuanced understanding of both the interactive scene and textual instructions.

We aspire to endow digital humans with capabilities for *grounding* and *rearrangement*, effectively merging scenes

*These authors contributed equally.
†Cewu Lu is the corresponding author.

and instructions to generate fluid and natural rearrangement actions. To this end, we introduce a dataset of **F**ull-body **A**R-driven **V**irtual **O**bject **R**earrangement (**FAVOR**). Utilizing this large-scale dataset, we also develop a pipeline for motion rearrangement generation of digital human guided by **I**nstruction **Te**xt (**FAVORITE**).

For collecting **FAVOR** dataset, we develop a data collection platform (see Fig. 2) that integrates a motion capture (MoCap) system for recording body-hand motion with an AR system. This AR system furnishes participants with real-time feedback on the scene-body interactions, allowing the subjects to instantly adjust their motion based on the AR projection. During the recording of each motion rearrangement, a randomly generated, realistic scene is projected onto the screens in AR glasses. The subject is then requested to rearrange the scene objects based on a text instruction such as 'put the wine bottle on the top of the rack' (see Fig. 1). Reflective markers attached to the subject allow the real-time capturing of the body and hand poses using the MoCap system. The resulting hand mesh is projected into the AR glasses for reference. In total, we record 3k rearrangement sequences - equivalent to 7.17M frames of interaction data. Empirical evidence shows that employing an AR-integrated system, as opposed to using real objects for each scene, greatly enhances the efficiency of data collection, with only a slight compromise in accuracy. We additionally employ a variety of pre- and post-processing strategies to ensure a high level of realism and diversity.

Utilizing the **FAVOR** dataset, we also develop a pipeline for text-guided motion rearrangement generation. When presented with a complex scene and textual instructions, our goal is to parse the command, comprehend the scene, and produce a motion sequence for an avatar. We name the pipeline as **FAVORITE** (**ITE** stands for **I**nstruction **TE**xt). To achieve this, we establish a practical framework with two stages: 1) Scene-Language Grounding and 2) Motion Rearrangement Generation. In the first stage, we initially transform the unstructured text instructions into Python-like code with a template function: `locate`. Then, we render the scene to multi-view observations and exploit a vision-language object detector (Minderer et al. 2022) to detect the object-of-interest in each view. Finally, we back-project object location from images to the 3D scene via multi-view geometry. This allows us to identify the object's initial and final positions and orientations. In the second stage, we synthesize the rearrangement motion in two steps. First, with both the initial and anticipated object location, we generate the two full-body grasping poses as *keyframes*. These keyframe poses are yielded from a CVAE. Then, we exploit an in-betweening module to yield a coherent and natural motion sequence between those keyframe poses.

To summarize, our study contributes the following: we propose a large-scale human rearrangement dataset, **FAVOR**, that encompasses an abundance of action sequences while demonstrating extensive scalability. We also design **FAVORITE** that can generate accurate, natural rearrangement sequences, by parsing instructional texts.

Related Works

Human-Object Interaction Datasets Human-object interaction is a critical aspect of study in CG, AR/VR, and robotics. It holds relevance for understanding human behavior, scene comprehension, and the advancement of embodied intelligence. Most existing datasets principally focus only on body poses, as illustrated by Mahmood et al. (2019), where a large and diverse human motion database, AMASS, is proposed. Another work (Cai et al. 2022) presents a multi-modal 4D human dataset for versatile sensing and modeling. However, these neglect facial expressions, gestures, and precise posture descriptions, which limit their usability in human-object interaction situations. Several datasets specifically target hand-object interactions. Some are synthesized via rendering techniques (Hasson et al. 2019; Gao et al. 2022; Li et al. 2023), whereas others are collected through the annotation of real-world video captures (Chao et al. 2021; Yang et al. 2022b; Liu et al. 2022). Despite ensuring accuracy in hand-object poses, these datasets overlook the coordination of whole-body movements.

In contrast to datasets that include hand-object interaction, there are relatively fewer datasets on whole-body and object interaction (Bhatnagar et al. 2022; Zhang et al. 2023). The paucity is mainly due to the complexity and time-consuming nature of collecting and labeling full-body and object interaction datasets. Very recently, Araújo et al. (2023) utilizes VR equipment to collect the CIRCLE dataset. However, the dataset still lacks hand pose annotations. The dataset that is close to ours is GRAB (Taheri et al. 2020), which provides rich annotation information, including whole-body and hand-object interactions with intent. However, the shortcoming of GRAB is that it only has homogeneous scenarios with almost invariable object positions. Our **FAVOR** dataset focuses on human rearrangement motions, a basic and crucial part of the manipulation. The diverse scenarios in **FAVOR** are instrumental in promoting human behavior, scene understanding, and embodied AI.

Motion Synthesis The target of motion synthesis is to create believable and fluid human action sequences. A majority of works utilize stochastic models to synthesize human motions (Ling et al. 2020; Liu et al. 2021) or human interaction with scenes (Wang et al. 2021, 2022). Lately, there has been an increase in works integrating language for action generation (Guo et al. 2020; Petrovich, Black, and Varol 2021). Petrovich, Black, and Varol (2022) employed a VAE-based Transformer module in TEMOS, while Tevet et al. (2022) harnessed a motion diffusion model in MDM to synthesize movements from textual descriptions. MotionGPT (Jiang et al. 2023) aligned the motion and language in a unified latent space.

Considering human-object interaction, generating motions becomes significantly more challenging due to the physical constraints. Some works generate single-frame grabs (Karunratanakul et al. 2020; Jiang et al. 2021; Yang et al. 2022a; Tendulkar, Surís, and Vondrick 2023), while others learn the implicit representation of motion to recover basic human actions in generating interactive sequences. For example, Taheri et al. (2022) used a conditional VAE to cre-

ate a 'goal' full-body grip and devised motions in an autoregressive manner. Wu et al. (2022) introduced a multi-task generative model that enables joint learning of static whole-body gripping pose and human-object contact.

These provided groundwork only for sequence generation involving grabbing. Notably, the concurrent works, IMoS and TOHO (Ghosh et al. 2023; Li et al. 2024) go further by attempting to create manipulation motions involving moving objects. However, these methods sidestepped complex scenarios, and the generated action sequences show room for improvement in terms of detail accuracy and fluency. Our work aims to overcome the setbacks of previous methods and accomplish the grasping and rearrangement of objects in complex scenes. We also employ a large language model to inject more contextual information and prior knowledge, such as scene semantics, physical constraints, etc., to improve the capability of modeling complex scenarios and character interactions.

FAVOR Dataset

Hardware Setup

To construct the dataset, we develop a platform that integrates AR glasses and a motion capture system. The infrared motion capture system includes 12 temporally synchronized Optitrack Prime 13W infrared cameras used for tracking reflective markers (Fig. 2 I.). We instruct the subject to wear a snug bodysuit outfitted with 18 markers to capture their full-body pose. To track hand movements, we affix 7mm markers directly on the subjects' skin using double-sided tape. (Fig. 2 II.) We utilize the XREAL X AR glasses for scene rendering. To track the AR glasses in the MoCap system, we attached markers to them and aligned the virtual coordinates from the glasses with real-world coordinates. (Fig. 2 III.)

Capture Protocol

We briefly elaborate on some of our data collection protocols. For more details, please refer to the Appx.

Pre-recording Approaches Before initiating the official recording, we project the scene that aligns with the real world onto the AR glasses. Additionally, we set up two virtual screens within the glasses to display left and right perspectives, improving users' precision in object manipulation. We also display textual instructions on the screen, such as, 'put the bottle on top of the white microwave.' Please note that due to real-time constraints, we only render several simplified geometric primitives (e.g., the white box) during data gathering. We substitute these primitives with more complex objects to diversify the dataset after data collection. Participants are asked to rehearse the rearrangement actions pre-recording to ensure smooth and natural motion during the recording phase.

Recording Object Movement During data collection, we mark a 'grasping' state when the subject's hand grasps the object by pressing a key on the keyboard. We then bind the global transformation $\mathcal{T}_{\mathcal{O}} \in \mathbb{R}^{4\times3}$ of the object $\mathcal{O}$ to the hand, allowing the object to move synchronized with the hand. When the object is placed, we note a 'release' state, thus ending the bond between the object and hand. Therefore, we can record the object's movement: $\mathcal{T}_{\mathcal{O}}^{0:T}$, where T is the length of the clip. This setting significantly conserves computational resources compared to physical simulation methods, facilitating real-time rendering. Based on our empirical observations, we found that it also assures better sequence continuity than automatic grasping (e.g. employing adhesion algorithms).

Figure 2: Illustration of recording setup (MoCap + AR glasses).

Motion Acquisition

We employ SMPL-X $\boldsymbol{\mathcal{S}}$, a parametric model for the human body, to portray an avatar's pose. The model $\boldsymbol{\mathcal{S}}$ is driven by the shape parameter $\boldsymbol{\beta} \in \mathbb{R}^{10}$, global translation $\boldsymbol{t} \in \mathbb{R}^{3}$, and the full-body pose $\boldsymbol{\theta} \in \mathbb{R}^{55\times3}$, which is represented in axis-angle. $\boldsymbol{\theta}$ includes the body parameter $\boldsymbol{\theta}_b$, the two hand parameters $\boldsymbol{\theta}_h$. We mark the posture parameters $\{\boldsymbol{\theta}, \boldsymbol{t}\}$ as $\boldsymbol{\Theta}$.

Without loss of generality, we only record operations for the right hand; operations for the left hand can be achieved using mirrored flip data. In the following sections, $\boldsymbol{\theta}_h$ represents only the pose of the right hand. These parameters enable us to compute the human body's mesh vertices $\mathcal{V} \in \mathbb{R}^{10475\times3}$ in a differentiable manner, with right-hand mesh vertices $\mathcal{V}_h \subset \mathcal{V}$, and $\mathcal{V}_h \in \mathbb{R}^{778\times3}$.

To guarantee real-time data capture and result precision, we only render the right-hand mesh $\mathcal{V}_h$ in the VR glasses. We track hand markers $\mathcal{M}_h \in \mathbb{R}^{8\times3}$ through motion capture (Fig. 2 II.), and caculate the hand pose $\boldsymbol{\theta}_h$ by resolving the inverse kinematic (IK). To better align the rendering results with actual conditions, we pre-scan the hand shape $\boldsymbol{\beta}_h$ of each subject.

After data collection, we restore the pose $\boldsymbol{\theta}_b$ with the collected reflective markers $\mathcal{M}_b \in \mathbb{R}^{17\times3}$ affixed to the full body through an optimization process. Specifically, we predefine 17 key points M_b on the human mesh $\mathcal{V}$, based on the snug bodysuit attached with markers (Fig. 2 II.). Thus, we can obtain the most optimal $\boldsymbol{\theta}_b$, $\boldsymbol{\beta}$, and $\boldsymbol{t}$, the result of minimizing the distance between M_b and $\mathcal{M}_b$.

Level of Realism

In our dataset, **FAVOR**, we construct interactive scenarios within a virtual space, with subjects engaging these environments via VR glasses. Subsequently, we ponder: How can we make motions that interact with virtual objects as realistic as with real objects? To address this challenge, we design

Dataset	#clip	#frame†	#fps	#$\mathcal{O}$ mesh	text / intent	3D anno.	body pose	hand pose	anno.
HAKE‡ (Li et al. 2022)	0.3 K	0.23 M	1	–	✓	✗	✗	✗	crowd
AMASS (Mahmood et al. 2019)	26 K	3.5 M	–	–	✗	✓	✓	✓	mix
HO3D (Hampali et al. 2020)	0.07 K	18.4 K	30	10	✗	✓	✗	✓	auto
ContactPose (Brahmbhatt et al. 2020)	2.3 K	0.99 M	–	25	✗	✓	✗	✓	auto
DexYCB (Chao et al. 2021)	1 K	72.5 K	30	20	✗	✓	✗	✓	crowd
OakInk‡ (Yang et al. 2022b)	1 K	57.5 K	30	100	✓	✓	✗	✓	crowd
HOI4D (Liu et al. 2022)	4 K	2.4 M	15	800	✓	✓	✗	✓	crowd
BEHAVE (Bhatnagar et al. 2022)	0.32 K	3.8 K	30	20	✗	✓	✓	✓	auto
HODome (Zhang et al. 2023)	0.27 K	0.93 M	60	23	✗	✓	✓	✓	auto + MoCap
GRAB (Taheri et al. 2020)	1.3 K	1.62 M	120	51	✓	✓	✓	✓	MoCap
FAVOR (Ours)	3 K	7.17 M	120	1800	✓	✓	✓	✓	AR + MoCap

Table 1: Statistics of the current human motion datasets. †: For the multi-view datasets, we only calculate the total number of frames within a single viewpoint. ‡: We only consider the video clips in the dataset. Annotation methods range from 'crowd': labeled by humans; 'auto': annotations from visual cues like segmentation, pose estimation, etc.; or 'mix': collected from multiple datasets.

the following strategies:

Physics Constraint To ensure that interactions echo real-world behaviors, we introduce a physics simulator. As a first step, to simulate the arbitrary positioning of objects, we generate random scenes and then allow objects to spontaneously settle on a virtual table. After participants finish placing objects, we rely on the physics simulator to confirm whether the positioning of virtual entities remains stable or if any penetration has occurred between them.

Body Prior To enhance the fidelity of body posture fitting result, we project the comprehensive body posture parameter $\mathbf{\Theta}$ into the Vposer latent space (Pavlakos et al. 2019). This integration introduces a static pose prior knowledge of human posture to $\mathbf{\Theta}$, thereby making the avatar more authentic and lifelike. Furthermore, we employ HuMoR (Rempe et al. 2021), a prior for human motion, to improve the continuity and fluidity of our temporal data.

Optimize Grasp VR settings often provide visual cues for grasping objects. Unfortunately, they fall short of providing tactile information. We need to mitigate incidents of hands and objects either penetrating each other or not making contact. Specifically, we train a network (Park et al. 2019) for each object $\mathcal{O}$ to calculate the SDF value. The function $SDF_{\mathcal{O}}(x)$ returns the minimal distance from the query point x to the object's mesh surface. If $SDF_{\mathcal{O}}(x) > 0$, it suggests that the point is outside the mesh, and vice versa. Eq. (1) penalizes all hand vertices that are within the object.

$$L_{\text{penetrate}} = \sum_{i} -\min(SDF_{\mathcal{O}}(\mathcal{V}_{h,i}), 0) \tag{1}$$

To address unstable contact, we draw the anchors (which are placed on the hand surface) toward their corresponding contact regions on the target object. The anchors are attached to 17 hand regions $\mathcal{A} = \{\mathcal{A}_i\}_i^{17}$ from the hand mesh $\mathcal{V}_h$ similar with Yang et al. (2021). The indicator function $\mathcal{I}_{j,k}$ represent whether the distance between anchor $\mathcal{A}_j$ and object vertex $\mathcal{V}_{\mathcal{O},k}$ is less than 3 cm, the overall loss is Eq. (2)

$$L_{\text{contact}} = \frac{1}{\sum \mathcal{I}} \sum_{j,k} \mathcal{I}_{j,k} ||\mathcal{A}_j - \mathcal{V}_{\mathcal{O},k}||_2^2 \tag{2}$$

Thus, the total optimization objective can be expressed as:

$$\underset{\boldsymbol{\theta}_h, \mathcal{T}_{\mathcal{O}}}{\operatorname{argmin}} \left(L_{\text{penetrate}} + L_{\text{contact}}\right) \tag{3}$$

From a perceptual study, we find that the sequences portraying interactions with virtual objects in our **FAVOR** dataset are strikingly similar to the way interactions with real-world objects.

Following the post-processing of the action sequences, they are distributed amongst a group of five volunteers for perceptual assessment. In this evaluation, the volunteers are required to evaluate the naturalness of the body movements, the appropriateness of the hand's grip on the object, and the consistency of the sequence with physics limitations. Only those sequences that reach a unanimous consensus on believability amongst the five volunteers are selected for further consideration.

Level of Diversity

To enrich and diversify **FAVOR**, we design a multi-step data augmentation process. Initially, we conceive diverse scenarios including but not limited to kitchen environments, office desks, laboratory benches, and coffee tables. In these scenarios, simple primitives as mentioned before are replaced with the original textured and complex objects borrowed from OmniObject3D (Wu et al. 2023). Enhancing the data further, we leveraged GPT-4 (OpenAI 2023), a powerful language model, to transform the given inflexible instruction 'put the bowl on top of the oven' into a naturally worded directive: 'Please place the bowl carefully on top of the microwave oven.' Simultaneously, we incorporated an additional level of complexity to each scenario by voxelizing the scene, identifying non-intersecting spaces within the human movement trajectory, and then selectively populating these spaces with random objects to add challenge. Details of the enrichment method are described in Appx.

Dataset Analysis

In conclusion, our **FAVOR** dataset contains 3k unique rearrangement sequences with diverse scenes with a variety of objects and instructions. We compare **FAVOR** with popular human motion datasets in Tab. 1. **FAVOR** incorporates 1,800 various object models allocated for rearrangement. We also

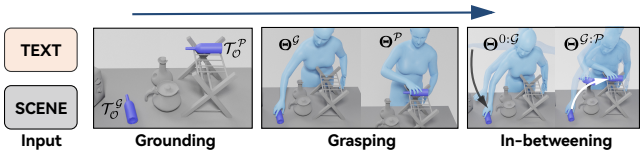

Figure 3: FAVORITE: text-guided motion rearrangement pipeline. It sequentially grounds the object locations, generates grasp poses and fills the motions in between.

employ the textured objects in OmniObject3D (Wu et al. 2023) to enrich the scene diversity. Approximately, each scene possesses 2-8 objects placed at random. We manually design approximately half of the commands to express the concept of 'be placed on top of', an operation that necessitates an understanding of object placement stability to prevent objects from rolling off—a task that presents a significant challenge for avatars to learn.

Each sequence records about 20 seconds, amounting to around 2,400 frames. Thus, we have collectively gathered 7.17 million motion data. In terms of traditional MoCap recording procedures, scene setup consumes 3 minutes, object rearrangement requires 1 minute, and post-processing operations (consisting of ghost point removal, marker relabeling, result fitting, etc.) claim about 45 minutes. However, in **FAVOR**, with the implementation of AR, post-processing is nearly eliminated due to minimal obstruction, reducing the time taken to achieve a sequence to just 3 minutes. This increased efficiency enables us to boost data collection rates by 15 times.

FAVORITE Pipeline

Based on the **FAVOR** dataset, we also develop a pipeline for text-guided motion rearrangement generation, which we call **FAVORITE**, where **ITE** stands for **I**nstruction **Te**xt. Given a 3D scene and the accompanying instructions, our goal is to synthesize motion for an avatar such that it rearranges a specified object in accordance with the instructions. This task is divided into two main components: 1) **Scene-Language Grounding**, which involves parsing the instructional text in conjunction with the scene to establish the object's initial and target final locations; 2) **Motion Rearrangement Generation**, which entails synthesizing a sequence of movements for the avatar to adeptly rearrange the object within the confines dictated by the scene. The workflow of our approach is depicted in Fig. 3.

Scene-Language Grounding

The key to the grounding problem is to figure out 1) $\mathcal{T}_{\mathcal{O}}^{\mathcal{G}}$: where the object is so that the avatar can **G**rasp it, and 2) $\mathcal{T}_{\mathcal{O}}^{\mathcal{P}}$: where the object should be **P**laced. Similar to the strategy used in VoxPoser (Huang et al. 2023), we utilize off-the-shelf solutions to solve these. Firstly, we harness the capabilities of a large language model, specifically GPT-4 (OpenAI 2023), to parse the instruction to `objects` and `preposition`, including the object to be grasped $\mathcal{O}$, along with the placement preposition (e.g. 'on top of the rack'). Next, to locate the object's initial position and to anticipate its final position from the scene, we design a template function: `locate`, which takes the scene information and textual instruction as input and outputs the spatial coordinates of the object. Considering real-world complexities where acquiring comprehensive scene structures can prove difficult, our method leverages a multi-view representation of the scene. Therefore, to achieve `locate`, we first position virtual cameras across the scene and render it into high-quality images from multiple viewpoints. Then we use the state-of-the-art object detection model, Owl-ViT (Minderer et al. 2022), to identify the `object`'s 2D bounding box in the images. Finally, we use the multi-view geometry to lift the 2D positions to a 3D bounding box and retrieve the `object`'s initial location $\mathcal{T}_{\mathcal{O}}^{\mathcal{G}}$ and anticipated location $\mathcal{T}_{\mathcal{O}}^{\mathcal{P}}$. The process is illustrated in Fig. 4.

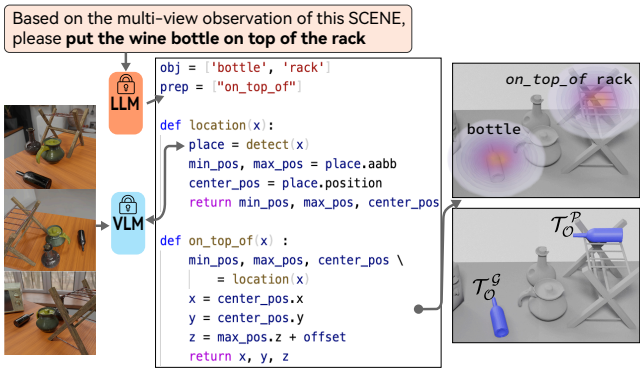

Figure 4: Diagram of visual-language grounding procedure. Text and images are parsed through the LLM and VLM to implement `locate` function, and the outcomes of both initial and anticipated object locations ($\mathcal{T}_{\mathcal{O}}^{\mathcal{G}}, \mathcal{T}_{\mathcal{O}}^{\mathcal{P}}$) are subsequently established.

Motion Rearrangement Generation

Considering the challenges caused by error accumulation when synthesizing entire sequences of human-object interactions in an auto-regressive manner, we propose a multi-stage approach to dividing the interaction process. As illustrated in Fig. 5, for the frame $f \in [0, T]$, we identify two critical frames: **the grasp frame** $\mathcal{G}$ and **the place frame** $\mathcal{P}$. We then generate the keyframe static full body grasping poses $\boldsymbol{\Theta}^{\mathcal{G}}$ and $\boldsymbol{\Theta}^{\mathcal{P}}$, corresponding to the objects' locations $\mathcal{T}_{\mathcal{O}}^{\mathcal{G}}$ and $\mathcal{T}_{\mathcal{O}}^{\mathcal{P}}$, respectively. Subsequently, we focus on generating the motion sequence $\boldsymbol{\Theta}^{0:\mathcal{G}}$, which transitions from the human's zero T-pose $\boldsymbol{\Theta}^{0}$ to the grasp pose $\boldsymbol{\Theta}^{\mathcal{G}}$, leveraging techniques similar to those employed in GOAL and SAGA (Taheri et al. 2022; Wu et al. 2022). After that, we take the step further to generate the rearrangement motion $\boldsymbol{\Theta}^{\mathcal{G}:\mathcal{P}}$, incorporating $\boldsymbol{\Theta}^{\mathcal{G}}$ and $\boldsymbol{\Theta}^{\mathcal{P}}$ as keyframe reference poses.

Keyframe Grasp Pose Generation Given an object $\mathcal{O}$ and its 6D pose $\mathcal{T}_{\mathcal{O}}$ within the scene, we design a **K**eyframe generation **NET**work, **KNET**, to create static full-body pose parameters $\boldsymbol{\Theta}$ that stably grasp the object. To extract the shape information of the object more efficiently, we utilize Basis Point Set (BPS) (Prokudin, Lassner, and Romero 2019) $\boldsymbol{b}_{\mathcal{O}} \in \mathbb{R}^{1024}$ as our representational form. Consistent with GNet in GOAL (Taheri et al. 2022), **KNET** is a conditional variational auto-encoder (cVAE). During the train-

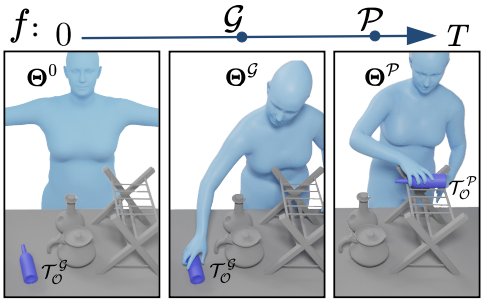

Figure 5: Diagram of motion rearrangement generation. keyframe grasping poses ($\mathbf{\Theta}^{\mathcal{G}}$, $\mathbf{\Theta}^{\mathcal{P}}$) are firstly generated and the sequence is full-filled via inbewteening ($\mathbf{\Theta}^{0:\mathcal{G}}$, $\mathbf{\Theta}^{\mathcal{G}:\mathcal{P}}$).

ing phase, the data relating to body $[\mathbf{\Theta}, \boldsymbol{\beta}, \boldsymbol{v}, \boldsymbol{d}_{b\to\mathcal{O}}]$ and object information $[\boldsymbol{b}_{\mathcal{O}}, \mathcal{T}_{\mathcal{O}}]$ are passed to the **KNET**'s encoder. The parameter $\boldsymbol{v} \in \mathbb{R}^{400\times3}$ represents vertices sampled from the body mesh $\mathcal{V}$, with a higher probability of sampling points closer to the object $\mathcal{O}$. $\boldsymbol{d}_{b\to\mathcal{O}} \in \mathbb{R}^{N\times3}$ is a set of the offset vectors from the sampled body vertices $\boldsymbol{v}$ to their closest objects vertices $\mathcal{V}_{\mathcal{O}}$. The encoder then maps all the input data into normal distribution parameters $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ via the re-parameterization trick. The decoder of **KNET** then uses $[\boldsymbol{b}_{\mathcal{O}}, \mathcal{T}_{\mathcal{O}}]$ and a sampled variable $\boldsymbol{z} \in \mathbb{R}^{16}, \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$ to predict $[\hat{\mathbf{\Theta}}, \hat{\boldsymbol{v}}_h, \hat{\boldsymbol{d}}_{h\to\mathcal{O}}]$. In this context, $\boldsymbol{v}_h \subset \mathcal{V}_h$ refers to the sampled hand vertices, and $\boldsymbol{d}_{h\to\mathcal{O}}$ defines a group of offset vectors from $\boldsymbol{v}_h$ to $\mathcal{V}_{\mathcal{O}}$. We employ L1 losses to drive all the predicted parameters (the hat symbol $\hat{\cdot}$) towards the ground truth values (the non-hat symbol $\cdot$). At the same time, the Kullback-Leibler divergence or KL loss is used to align the $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$ closer to $\mathcal{N}(\mathbf{0}, \mathbf{1})$.

Since the human pose $\mathbf{\Theta}$ has a high degree of freedom, the network-predicted pose $\hat{\mathbf{\Theta}}$ often encounters issues like hand-object penetration or unstable grasping. To address this, we leverage the losses defined in Eq. (3) and implement a regularization loss $L_{\text{reg}} = ||\mathring{\mathbf{\Theta}} - \hat{\mathbf{\Theta}}||_1$, which helps prevent the optimized pose $\mathring{\mathbf{\Theta}}$ from deviating from the predicted pose $\hat{\mathbf{\Theta}}$. Our overall goal is expressed in Eq. (4). Now with the object's initial pose $\mathcal{T}_{\mathcal{O}}^{\mathcal{G}}$ and final pose $\mathcal{T}_{\mathcal{O}}^{\mathcal{P}}$, we can infer the optimized grasp body pose $\mathring{\mathbf{\Theta}}^{\mathcal{G}}$ and $\mathring{\mathbf{\Theta}}^{\mathcal{P}}$, respectively.

$$\underset{\mathring{\mathbf{\Theta}}}{\operatorname{argmin}} \left(L_{\text{penetrate}} + L_{\text{contact}} + L_{\text{reg}}\right) \tag{4}$$

In-Betweening Motion Generation With two keyframe grasp poses $\mathbf{\Theta}^{k_1}$ and $\mathbf{\Theta}^{k_2}$, we establish an **I**nbetweening **NET**work, **INET**, to generate human motion $\mathbf{\Theta}^{k_1:k_2}$. **INET** is accomplished via modifications to an auto-regressive based MNet (Taheri et al. 2022). When considering the current frame t, where $t \in [0, T]$, we take $[\mathbf{\Theta}^{t-5:t}, \boldsymbol{\beta}, \boldsymbol{v}^t, \boldsymbol{v'}^t, \boldsymbol{d}_{h^t\to h^k}, \boldsymbol{b}_h^k], k \in \{k_1, k_2\}$ as input. Here, $\mathbf{\Theta}^{t-5:t}$ represents the body pose from the previous 5 frames. $\boldsymbol{v'}^t$ is the current frame's velocity of body vertices. $\boldsymbol{d}_{h^t\to h^k}$ represents a series of offset vectors of hand vertices from the current frame t to the keyframe k, and $\boldsymbol{b}_h^k$ is the BPS representation of hand vertices at keyframe k. To reduce the complexity of the predictions, the output of **INET** presents as the changes, denoted by $\Delta\cdot$, associated with the current frame over the future 10 frames: $[\Delta\hat{\mathbf{\Theta}}^{t:t+10}, \Delta\hat{\boldsymbol{t}}^{t:t+10}, \Delta\hat{\boldsymbol{v}}^{t:t+10}, \Delta\hat{\boldsymbol{d}}_{h^{t:t+10}\to h^k}]$. During the inference stage, **INET** iteratively predicts the sequence of motion - that is, it uses $\hat{\mathbf{\Theta}}^{t+10} = \Delta\hat{\mathbf{\Theta}}^{t+10} + \hat{\mathbf{\Theta}}^t$ as the input for the next cycle, until the hand approaches the object target position $\mathcal{T}_{\mathcal{O}}^{k_2}$.

In striving to minimize error accumulation arising from the auto-regressive manner, we propose a two-step approach to motion synthesis: Initially, we train **INET** with $k_1 = 0, k_2 = \mathcal{G}$, which generates motion from the T-pose to the grasp pose $\mathbf{\Theta}^{0:\mathcal{G}}$. Subsequently, we train another **INET** to generate motion that progresses towards the placement pose, denoted as $\mathbf{\Theta}^{\mathcal{G}:\mathcal{P}}$, with $k_1 = \mathcal{G}, k_2 = \mathcal{P}$.

Obstacle Avoidance Algorithm Given the challenge of encoding obstacle data into the motion synthesis network, we have designed an obstacle avoidance algorithm. Our design begins by simplifying all N scene obstacles, comprising tables, into an enumerated set of shape primitives: $\{\mathcal{Q}_i | i \in [0, N]\}$. We then efficiently calculate the Signed Distance Function (SDF) for each primitive as $SDF_{\mathcal{Q}_i}$. For every auto-regressive prediction of $\hat{\mathbf{\Theta}}^t$ made by **INET**, we calculate the body vertices with SMPL-X: $\hat{\mathcal{V}} = \mathcal{S}(\hat{\mathbf{\Theta}}^t, \hat{\boldsymbol{t}}^t)$. Leveraging Eq. (5), we can push out the vertices within these obstacles and get the optimized $\mathring{\mathbf{\Theta}}^t$. Then we treat $\mathring{\mathbf{\Theta}}^t$ as the input frame pose to the next iteration of **INET**. We also add a regularization item to make sure the optimized pose $\mathring{\mathbf{\Theta}}^t$ is close to the predicted pose $\hat{\mathbf{\Theta}}^t$.

$$\underset{\mathring{\mathbf{\Theta}}^t}{\operatorname{argmin}} \sum_{i,j} -\min(SDF_{\mathcal{Q}_i}(\hat{\mathcal{V}}_j), 0) + ||\mathring{\mathbf{\Theta}}^t - \hat{\mathbf{\Theta}}^t||_1 \tag{5}$$

Object Motion While our study primarily focuses on predicting the movements of digital humans, a comprehensive analysis cannot disregard the associated trajectories of moving objects, denoted as $\hat{\mathcal{T}}_{\mathcal{O}}^t, t \in [0, T]$. Establishing a specialized network for this task is not just expensive, but it also results in an unstable grasp. During the rearrangement process, an object remains unvaryingly gripped by the hand. Taking this into account, during the rearrangement frame from $\mathcal{G}$ to $\mathcal{P}$, we bind the pose of the object $\mathcal{O}$ with the hand. In other words, we fix the transformation of the object relative to the hand, denoted as $\mathcal{T}_{h\to\mathcal{O}}^t$. Aware of minor finger movement that leads to penetration, we transfer each hand vertices $\mathcal{V}_{h,j}, j \in [1, 778]$ to the object's coordinate. We minimize penetration by treating Eq. (6) as the optimization target for each frame t.

$$\underset{\mathring{\mathcal{T}}_{h\to\mathcal{O}}}{\operatorname{argmin}} \sum_{j} -\min(SDF_{\mathcal{O}}(\mathring{\mathcal{T}}_{h\to\mathcal{O}}^{-1} \cdot \mathcal{V}_{h,j}), 0) \tag{6}$$

Experiments

Setup and Metrics

Setup We evaluate the **FAVOR** dataset and pipeline in terms of quality and performance. The dataset is split into training, validation, and test sets in an 8:1:1 ratio, based on motion sequences. Training details are in the Appx. During inference, objects in the scene are simplified to primitives for efficient collision detection analysis, as depicted in Fig. 6.

Perceptual Study We require 12 participants to assess both ground-truth and generated sequences, rating their quality

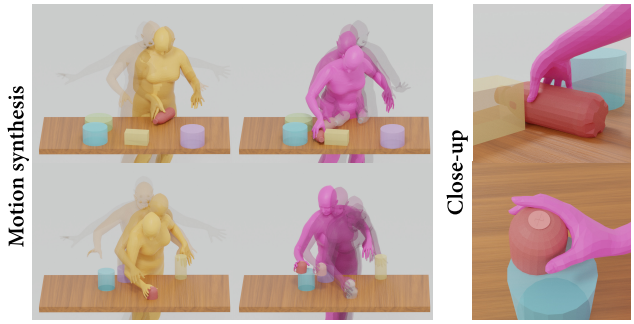

Figure 6: Qualitative results of motion synthesis. The man highlighted in yellow illustrates the grasping motion $\mathbf{\Theta}^{0:\mathcal{G}}$, and the man marked in magenta represents the placement motion $\mathbf{\Theta}^{\mathcal{G}:\mathcal{P}}$.

Metric	INET	FAVOR
Objective Completion	4.01 ± 1.02	4.67 ± 0.55
Physical Plausibility	3.64 ± 1.02	4.63 ± 0.53
Interaction Stability	3.75 ± 1.12	4.56 ± 0.53
Motion Naturalness	3.99 ± 0.98	4.63 ± 0.55
Average	3.85 ± 1.04	4.62 ± 0.54

Table 2: Perceptual Study of sequences in both generated motion and FAVOR dataset. The results are the average of Likert scores.

on a 1 (disagree) to 5 (agree) scale. The metrics include objective completion status, physical plausibility, interaction stability, and motion naturalness.

Evaluation Metrics We use three metrics to evaluate keyframe grasp pose generation: 1) Contact Ratio (CR) measures body-object contact within a 5mm threshold. 2) Solid Intersection Volume (SIV): We voxelize the objects and compute the volume of objects (measured in cm^3) that resides within the human body. 3) Average Pairwise Distance (APD): The diversity of both ground truth (**g.t.**) and predicted (**p.**) sequences are examined via the assessment of the average pairwise mean per-vertex position error (MPVPE, measured in m) of the human body surfaces.

Motion smoothness in the inbetweening module is evaluated using Power Spectrum KL divergence of joints (PSKL-J), similar to SAGA (Wu et al. 2022). This measures the acceleration distribution variance between predicted and **g.t.** joint sequences, reporting results in both directions.

Evaluation

Perceptual Study Results We present the perceptual study scores of **FAVOR** and our predicted motion in Tab. 2. Despite the dataset's origins in AR collection, it attains superior data quality through our pre-recording protocol and our post-processing algorithm. Additionally, the predicted motion exhibits exemplary performance in both completeness and naturalness.

	CR ↑	SIV ↓	APD ↑
FAVOR	0.99	0.83	0.36
KNET $\mathcal{G}$	0.99	7.05	0.17
KNET $\mathcal{P}$	0.96	7.61	0.18
KNET *all*	0.98	7.33	0.17

Table 3: Grasp quality evaluation for FAVOR and KNET.

Method	PSKL-J ↓ (g.t., p.)	PSKL-J ↓ (p., g.t.)
INET $0:\mathcal{G}$ w/ **g.t.**	0.237	0.193
INET $\mathcal{G}:\mathcal{P}$ w/ **g.t.**	0.491	0.511
INET $0:\mathcal{G}$ w/ **KNET p.**	0.255	0.209
INET $\mathcal{G}:\mathcal{P}$ w/ **KNET p.**	0.706	0.657

Table 4: PSKL-J score of our INET with different phase and keyframe grasping pose.

	CR ↑	SIV ↓
INET w/o obj opt	0.96	11.81
INET w/ *FLEX* **p.**	0.50	6.23
INET w/ **g.t.**	0.92	5.90
INET w/ **KNET p.**	0.89	6.50

Table 5: INET ablations. The SIV scores are averages from sequences grasping success.

Grasp Evaluation We present the quantification of grasping data from **FAVOR** and the associated **KNET** results in Tab. 3. **FAVOR** exhibits minor penetration errors and a high degree of pose variability. To evaluate the quality of **KNET**-generated grasps, we provide scores for grasped objects in both their initial $\mathcal{T}_{\mathcal{O}}^{\mathcal{G}}$ and final $\mathcal{T}_{\mathcal{O}}^{\mathcal{P}}$ poses, all calculated using ground truth (**g.t.**) objects pose to negate the effects of grounding. Furthermore, our findings suggest that the object poses for placement $\mathcal{T}_{\mathcal{O}}^{\mathcal{P}}$ result in less stable grasping results, implying that the task of placement may present more complexity than that of grasping.

Motion Evaluation The respective independent performances of **INET** $0:\mathcal{G}$ and **INET** $\mathcal{G}:\mathcal{P}$ are documented in Tab. 4. **g.t.** denotes that the ground truth keyframe body poses are provided to evaluate **INET**'s potential upper bound performance. Our findings indicate that the sequence **INET** $0:\mathcal{G}$, which originates from the T-pose $\mathbf{\Theta}^0$, outperforms **INET** $\mathcal{G}:\mathcal{P}$. This suggests that synthesizing rearrangements presents a greater challenge due to the increased constraints imposed by the actions.

Ablation As shown in the first row of Tab. 5, the SDF-based loss item defined in Eq. (6) can well alleviate the object penetration issue when grasped in the hand. We also explore using the grasping pose optimized from FLEX (Tendulkar, Surís, and Vondrick 2023) to substitute the **KNET** results as the keyframe pose. We find that since the FLEX has not been trained with rearrangement motions in desktop scenarios, the resulting motion is not as stable as the **KNET**, and nearly half of the attempts are unsuccessful.

Conclusion

In this paper, we present a comprehensive dataset **FAVOR**, created by combining MoCap and AR glasses. The dataset composes complex scenes, diversity instructions, and precise avatar rearrangement motions. Leveraging **FAVOR**, we build a pipeline, **FAVORITE**, to parse instructions and scenes, generate grasping poses, and fill natural motions in between. We anticipate that this dataset and pipeline will serve as a foundational bridge linking various tasks under the long-horizon and multi-task motion generation.

Acknowledgements

This work was supported by the National Key R&D Program of China (No. 2021ZD0110704), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), and Shanghai Science and Technology Commission (21511101200). Cewu Lu is the corresponding author of this work. He is the member of Qing Yuan Research Institute and MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, China.

References

Araújo, J. P.; Li, J.; Vetrivel, K.; Agarwal, R.; Wu, J.; Gopinath, D.; Clegg, A. W.; and Liu, K. 2023. CIRCLE: Capture In Rich Contextual Environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Bhatnagar, B. L.; Xie, X.; Petrov, I. A.; Sminchisescu, C.; Theobalt, C.; and Pons-Moll, G. 2022. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Brahmbhatt, S.; Tang, C.; Twigg, C. D.; Kemp, C. C.; and Hays, J. 2020. 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,*.

Cai, Z.; Ren, D.; Zeng, A.; Lin, Z.; Yu, T.; Wang, W.; Fan, X.; Gao, Y.; Yu, Y.; Pan, L.; Hong, F.; Zhang, M.; Loy, C. C.; Yang, L.; and Liu, Z. 2022. HuMMan: Multi-modal 4d human dataset for versatile sensing and modeling. In *17th European Conference on Computer Vision, Tel Aviv, Israel, October 23–27, 2022, Proceedings,*.

Chao, Y.-W.; Yang, W.; Xiang, Y.; Molchanov, P.; Handa, A.; Tremblay, J.; Narang, Y. S.; Van Wyk, K.; Iqbal, U.; Birchfield, S.; et al. 2021. DexYCB: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Gao, D.; Xiu, Y.; Li, K.; Yang, L.; Wang, F.; Zhang, P.; Zhang, B.; Lu, C.; and Tan, P. 2022. DART: Articulated hand model with diverse accessories and rich textures. *Advances in Neural Information Processing Systems*.

Ghosh, A.; Dabral, R.; Golyanik, V.; Theobalt, C.; and Slusallek, P. 2023. IMoS: Intent-Driven Full-Body Motion Synthesis for Human-Object Interactions. In *Computer Graphics Forum*, 2.

Guo, C.; Zuo, X.; Wang, S.; Zou, S.; Sun, Q.; Deng, A.; Gong, M.; and Cheng, L. 2020. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*.

Hampali, S.; Rad, M.; Oberweger, M.; and Lepetit, V. 2020. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Hasson, Y.; Varol, G.; Tzionas, D.; Kalevatykh, I.; Black, M. J.; Laptev, I.; and Schmid, C. 2019. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Huang, W.; Wang, C.; Zhang, R.; Li, Y.; Wu, J.; and Fei-Fei, L. 2023. VoxPoser: Composable 3D Value Maps for Robotic Manipulation with Language Models. *arXiv preprint arXiv:2307.05973*.

Jiang, B.; Chen, X.; Liu, W.; Yu, J.; Yu, G.; and Chen, T. 2023. MotionGPT: Human Motion as a Foreign Language. *arXiv preprint arXiv:2306.14795*.

Jiang, H.; Liu, S.; Wang, J.; and Wang, X. 2021. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Karunratanakul, K.; Yang, J.; Zhang, Y.; Black, M. J.; Muandet, K.; and Tang, S. 2020. Grasping field: Learning implicit representations for human grasps. In *2020 International Conference on 3D Vision (3DV)*.

Li, K.; Yang, L.; Zhen, H.; Lin, Z.; Zhan, X.; Zhong, L.; Xu, J.; Wu, K.; and Lu, C. 2023. CHORD: Category-level Hand-held Object Reconstruction via Shape Deformation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Li, Q.; Wang, J.; Loy, C. C.; and Dai, B. 2024. Task-oriented human-object interactions generation with implicit neural representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*.

Li, Y.-L.; Liu, X.; Wu, X.; Li, Y.; Qiu, Z.; Xu, L.; Xu, Y.; Fang, H.-S.; and Lu, C. 2022. Hake: a knowledge engine foundation for human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ling, H. Y.; Zinno, F.; Cheng, G.; and Van De Panne, M. 2020. Character controllers using motion vaes. *ACM Transactions on Graphics (TOG)*, (4).

Liu, Y.; Liu, Y.; Jiang, C.; Lyu, K.; Wan, W.; Shen, H.; Liang, B.; Fu, Z.; Wang, H.; and Yi, L. 2022. HOI4D: A 4D egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Liu, Z.; Lyu, K.; Wu, S.; Chen, H.; Hao, Y.; and Ji, S. 2021. Aggregated multi-gans for controlled 3d human motion prediction. In *Proceedings of the AAAI conference on artificial intelligence*, 3.

Mahmood, N.; Ghorbani, N.; Troje, N. F.; Pons-Moll, G.; and Black, M. J. 2019. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*.

Minderer, M.; Gritsenko, A.; Stone, A.; Neumann, M.; Weissenborn, D.; Dosovitskiy, A.; Mahendran, A.; Arnab, A.; Dehghani, M.; Shen, Z.; et al. 2022. Simple open-vocabulary object detection with vision transformers. *arXiv preprint arXiv:2205.06230*.

OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Park, J. J.; Florence, P.; Straub, J.; Newcombe, R.; and Lovegrove, S. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of*

the IEEE/CVF conference on computer vision and pattern recognition.

Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A.; Tzionas, D.; and Black, M. J. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Petrovich, M.; Black, M. J.; and Varol, G. 2021. Action-conditioned 3D human motion synthesis with transformer VAE. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Petrovich, M.; Black, M. J.; and Varol, G. 2022. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*.

Prokudin, S.; Lassner, C.; and Romero, J. 2019. Efficient learning on point clouds with basis point sets. In *Proceedings of the IEEE/CVF international conference on computer vision*.

Rempe, D.; Birdal, T.; Hertzmann, A.; Yang, J.; Sridhar, S.; and Guibas, L. J. 2021. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*.

Taheri, O.; Choutas, V.; Black, M. J.; and Tzionas, D. 2022. GOAL: Generating 4D whole-body motion for hand-object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Taheri, O.; Ghorbani, N.; Black, M. J.; and Tzionas, D. 2020. GRAB: A dataset of whole-body human grasping of objects. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings,*.

Tendulkar, P.; Surís, D.; and Vondrick, C. 2023. FLEX: Full-Body Grasping Without Full-Body Grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-Or, D.; and Bermano, A. H. 2022. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*.

Wang, J.; Rong, Y.; Liu, J.; Yan, S.; Lin, D.; and Dai, B. 2022. Towards diverse and natural scene-aware 3d human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Wang, J.; Xu, H.; Xu, J.; Liu, S.; and Wang, X. 2021. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Wu, T.; Zhang, J.; Fu, X.; Wang, Y.; Ren, J.; Pan, L.; Wu, W.; Yang, L.; Wang, J.; Qian, C.; et al. 2023. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 803–814.

Wu, Y.; Wang, J.; Zhang, Y.; Zhang, S.; Hilliges, O.; Yu, F.; and Tang, S. 2022. Saga: Stochastic whole-body grasping with contact. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings,*.

Yang, L.; Li, K.; Zhan, X.; Lv, J.; Xu, W.; Li, J.; and Lu, C. 2022a. ArtiBoost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2750–2760.

Yang, L.; Li, K.; Zhan, X.; Wu, F.; Xu, A.; Liu, L.; and Lu, C. 2022b. OakInk: A Large-scale Knowledge Repository for Understanding Hand-Object Interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yang, L.; Zhan, X.; Li, K.; Xu, W.; Li, J.; and Lu, C. 2021. CPF: Learning a contact potential field to model the hand-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Zhang, J.; Luo, H.; Yang, H.; Xu, X.; Wu, Q.; Shi, Y.; Yu, J.; Xu, L.; and Wang, J. 2023. NeuralDome: A Neural Modeling Pipeline on Multi-View Human-Object Interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.