

GigaHands: A Massive Annotated Dataset of Bimanual Hand Activities

Rao Fu^{1*} Dingxi Zhang^{2*} Alex Jiang¹ Wanjia Fu¹
Austin Funk¹ Daniel Ritchie¹ Srinath Sridhar^{1†}

* Equal contribution † Corresponding author

¹Brown University ²ETH Zurich

<https://ivl.cs.brown.edu/research/gigahands.html>

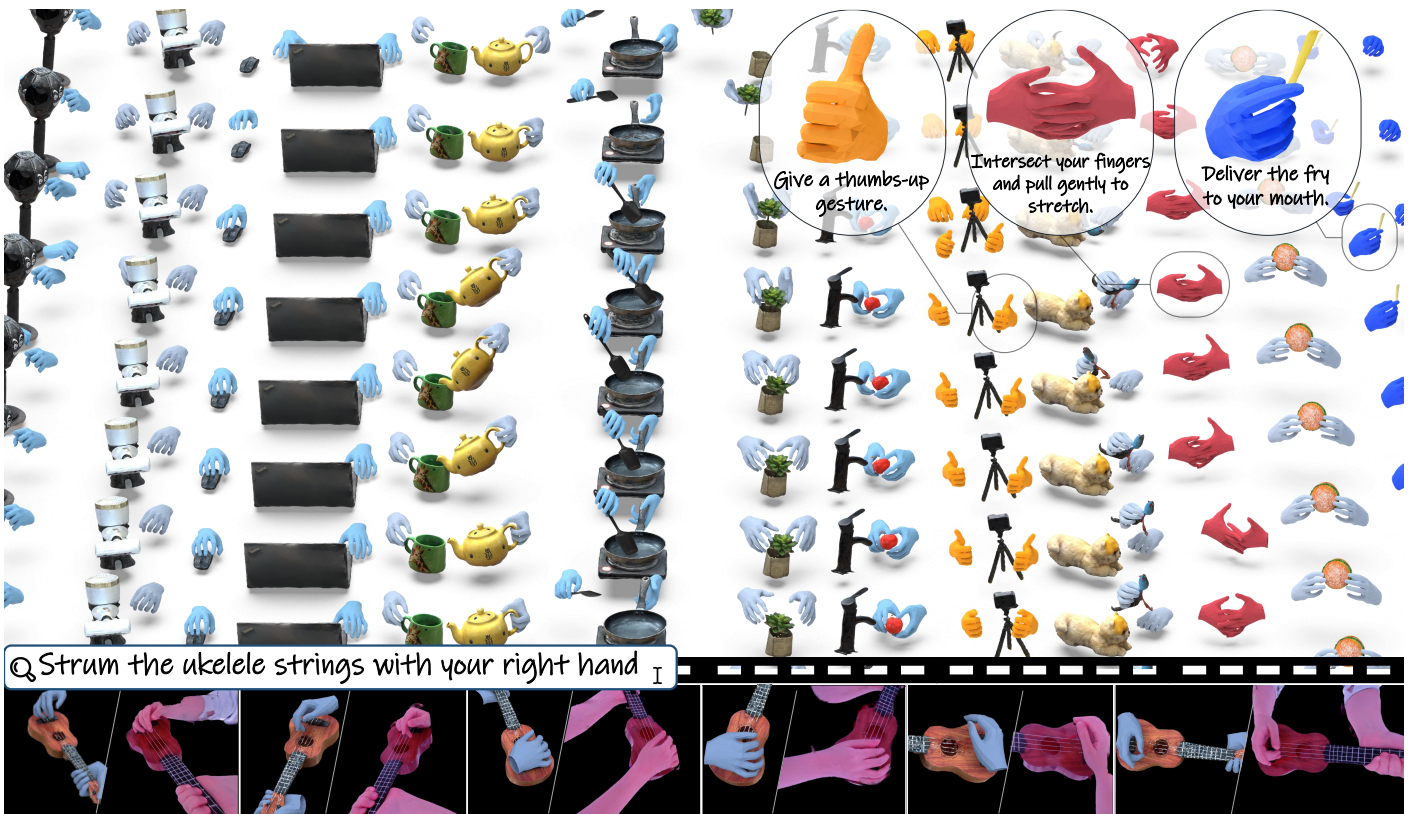


Figure 1. **GigaHands** is a massive dataset of human bimanual activities with paired text annotations. Each column above shows an activity sequence from the dataset. The dataset covers diverse 3D hand activities, including hand-object interactions (blue) across object scales, gestures (orange), and self-interactions (red). Each clip is paired with descriptive text and 51 camera views, enabling radiance field reconstruction. The bottom row show other annotations in the dataset including hand shape, object shape and pose (left half images). The right half images show novel views from dynamic radiance field fitting.

Abstract

Understanding bimanual human hand activities is a critical problem in AI and robotics. We cannot build large models of bimanual activities because existing datasets lack the scale, coverage of diverse hand activities, and detailed annotations. We introduce *GigaHands*, a massive annotated dataset capturing 34 hours of bimanual hand activities from 56 subjects and 417 objects, totaling 14k motion clips derived from 183 million frames paired with 84k text annotations.

Our markerless capture setup and data acquisition protocol enable fully automatic 3D hand and object estimation while minimizing the effort required for text annotation. The scale and diversity of *GigaHands* enable broad applications, including text-driven action synthesis, hand motion captioning, and dynamic radiance field reconstruction.

1. Introduction

“The **human hand** is a marvel of evolution, whose intricate structures and capabilities have allowed us to manipulate

the environment in ways no other species can.”

– F.R. Wilson [108]

From the skillful manipulation involved in cooking a meal, to the rapid movement of fingers to type this sentence, hands are always busy shaping our environments and communicating with others. Enabling machines to understand and replicate the remarkable capabilities of human hands is one of the grand challenges of AI and robotics. One approach to achieve this goal could be to train models on massive datasets, the strategy followed by recent large models that learn from trillions of text tokens [27], billions of text-image pairs [89], tens of millions of 3D shapes [25], or millions of robot trajectories [12, 13, 46, 73]. However, datasets of this scale are currently unavailable for natural hand activities.

Sourcing large-scale 3D hand activities data is challenging. The two most common methods for data acquisition involve using cameras to capture hand manipulations in the wild or in controlled studio settings. In-the-wild data includes monocular internet videos [91], egocentric videos captured using wearable cameras [22, 23, 32, 52, 57, 79], or multi-view videos from third-person cameras [30, 62, 91]. However, this data is sparse, hard to calibrate, and noisy, resulting in limited 3D motion reconstruction accuracy, especially for objects. Alternatively, studio settings bring subjects into camera-rich environments and could employ markers for accurate reconstruction. However, the staged setup and lack of real-world context limits data diversity, and marker-based tracking [8, 28, 29, 55, 58, 59, 101, 111, 116] inhibits natural interactions.

We address the data sourcing problem preventing the scaling up of hand datasets by introducing **GigaHands**, a diverse, massive, and fully-annotated 3D bimanual hand activities dataset (see Figure 1). To our knowledge, GigaHands is the largest bimanual hand activities dataset, with over **183 million** unique image frames with two hands each (0.37×10^9 unique hand poses, hence **GigaHands**). We used a multi-camera markerless capture system to acquire accurate bimanual hand-object interaction activities, gestures, and self-contacts. We overcome the studio capturing limitation by designing procedural activity elicitation protocols to include as many actions as in the real-world in-the-wild setting. To ensure diversity that reflects the real world, we captured data from **56 subjects** interacting with **417 real-world objects**. All images in our dataset are fully annotated with: detailed activity text descriptions; 3D hand shape and pose; MANO [86] hand meshes; 3D object shape, pose and appearance; hand/object segmentation masks; 2D/3D hand keypoints; camera pose. We minimized manual annotation effort using a procedural *instruct-to-annotate* strategy by guiding subjects with detailed instructions to reduce annotation effort post-capture. In total, we collected **34 hours** of text-annotated bimanual hand activities, **13k** 3D motion se-

quences, and over **14k** post-processed 3D motion clips with **84k** detailed atomic-level text description annotations. Our 3D hand motion clips exceed the combined size of all existing 3D bimanual hand activities datasets [57, 58, 116], and the range of verbs in our text annotations is larger than any other hand dataset, even those captured in the wild [32, 33].

GigaHands can unlock new capabilities in applications including motion generation, robotics, and dynamic 3D reconstruction [81]. To demonstrate, we show improvements enabled by our dataset on text-driven action synthesis and hand motion captioning, along with examples dynamic radiance field reconstruction. To sum up our contributions:

- We present **GigaHands**, a massive, diverse and annotated 3D bimanual activities dataset. It includes **34 hours** of activities, **14k** hand motions clips paired with **84k text annotation**, and over **183M** unique hand images.
- A **procedural data acquisition** strategy that ensures activity diversity and detailed atomic-level text descriptions while minimizing manual effort. We also built an accurate and fully-automated pipeline for estimating 3D hand/object shape and pose, segmentation, and camera pose.
- We show **applications** enabled by our dataset’s scale, including 3D hand motion generation, hand motion captioning, and dynamic semantic scene reconstruction.

2. Related Work

2.1. Hand Motion Data Sourcing

Hand motion data has four primary sources, and each presents its own advantages and challenges concerning realism, diversity, accuracy, and practicality.

Static poses involve motion planning [5, 7, 10, 93], policy learning [19, 60], or motion synthesis [20, 117, 118] from individual static positions. While these methods provide reliable 3D data, it lacks semantic richness, resulting in motions disconnected from real-world activities and not fully representing diverse human movement.

Synthetic data from game engines [126], augmented reality [66], virtual reality [38], or simulations [39, 63, 105] allow for controlled and accurate 3D motion sequences. Despite their reliability, these synthetic motions often fail to capture the intricacies of genuine human movements, limiting their applicability in modeling natural hand behaviors.

Real-world in-the-wild settings involves using wearable or portable sensors to capture hand interactions from egocentric [22, 23, 32, 52, 57, 79], third-person [30, 62, 91] or both types of views [33, 95], yielding realistic and natural motions. However, obtaining reliable 3D motion data for both hands and objects is challenging due to occlusions resulting from limited sensor deployment and accuracy. Additionally, annotating such data is labor-intensive, and certain self-contact motions are difficult to record accurately.

Studio environments place participants in controlled set-

tings equipped with extensive sensors, enabling reliable capture of detailed information such as 3D hand motions [45, 54, 55, 72, 120, 122], object movements [8, 18, 29, 36, 37, 48, 58, 59, 111, 116], contact regions [28, 80, 101], audio [44, 49, 71, 113], and tactile data [14, 31, 78, 100]. However, collecting data in studios is arduous and may not reflect the diversity of real-life scenarios. Participants might find it challenging to perform natural motions in an unfamiliar environment, especially when encumbered with wearable sensors. For example, motion capture (MoCap) systems [8, 28, 29, 55, 58, 59, 101, 111, 116] require markers attached to the body, reducing visual realism and inhibiting movements like self-contact between hands, necessitating additional post-processing to restore a realistic appearance. On the other hand, markerless motion capture systems [18, 36, 37, 45, 48, 54, 72, 80, 120, 122] offer more realistic visuals and encourage natural motions, but they may trade off capture accuracy. GigaHands balances accuracy, diversity, realism, and practicality by replicating in-the-wild settings during activity elicitation and estimating accurate 3D motion from marker-less motion capture.

2.2. Hand Datasets and Annotations

Hand datasets vary in sources and scales, each constructed for specific purposes and enriched with various annotations. **Hand motion** is the most common type of annotation, represented as bounding boxes [30, 79, 91], segmentation masks [24, 52, 70], 2D keypoints [6, 43, 61, 90], 3D keypoints [82, 99, 102, 106], or parameters of parametric models [55, 77, 97]. Depending on the data source and desired detail, annotations are obtained through manual labeling [106], synthesis [66, 67, 92, 126], markers and sensors [29, 45, 54, 104, 115, 120, 122], cross-view bootstrapping [65, 96], iteratively trained keypoint extraction networks [72, 127], multi-view annotation [72], or audio refinement [44]. Semi-automatic and automatic labeling techniques significantly improve scalability.

Object annotations are provided given the frequent interaction between hands and objects. These can be acquired from manual labeling [18, 57, 98], synthesis [21, 39], MoCap [8, 28, 58, 59, 111, 112, 116, 124], hand-object reconstruction or retrieval [15, 64, 109], or multi-view RGB-D data [36, 37, 48]. We demonstrate that multi-view RGB is sufficient for automatic annotation, enhancing scalability.

Text annotations. To capture the rich semantic meanings in human hand actions, certain datasets provide text annotations. These annotations may include action types [26, 30, 32, 47, 79], atomic action descriptions [22, 23, 33, 52, 90], activity narration [22, 23, 33, 62], activity commentary [17, 32, 33, 116], object affordance [42, 58, 111, 116], or body dynamics [114, 122, 123]. Typically, annotations are manually provided, especially when data are collected in uncontrolled, unscripted environments [22, 23, 32, 33].

Table 1. Comparisons of 3D bimanual motion datasets. Dataset names are highlighted with different colors if it has no text annotations (gray), **action type** (green), **sparse description** (red), and **dense description** (blue). See the supp. document for the full table.

Name	mins	motions	poses	views	frames	sub.	obj.
w/o Object Annotation							
AssemblyHands [72]	630	62	203k	12	3.03M	34	/
Ego-Exo4D [33]	/	/	4.4M	5-6	/	740	/
w/ Object Annotation							
HOI4D [57]	1,333	4k	1.2M	1	2.4M	4	800
ARCTIC [28]	121	339	218k	9	2.1M	10	11
TACO [58]	202	2.3k	363k	13	4.7M	14	196
OakInk2 [116]	557	2.8k	993k	4	4.01M	9	75
HOT3D [8]	833	4.1k	1.7M	2-3	3.7M	19	33
GigaHands (Ours)	2,034	13.9k	3.7M	51	183M	56	417

Even in studio settings, post-processing is often required to extract meaningful action clips [52, 116]. We introduce a procedural instruct-and-annotate pipeline that maximizes semantic interpretability and minimizes the annotation effort.

Contact are derived from manual annotation [91], video grounding [69], thermal sensors [9, 11] or geometry analysis [80, 101, 125] based on captured data. Our dataset enables us to derive contact regions using geometry analysis on both surface mesh and density fields.

3. The GigaHands Dataset

Dataset Characteristics. GigaHands is a large and diverse dataset of bimanual hand activities with detailed annotations. It encompasses a wide range of scenarios, including hand-object interactions, gesturing, and self-contacts performed by 56 subjects who collectively used 417 objects (see Table 1). The dataset contains 2,034 minutes of bimanual hand activities, surpassing the length of any existing 3D hand motion dataset. Compared to other datasets that are unannotated [8, 28], contain motion type annotations [57, 58, 72], or provide sparse motion descriptions [33, 116], GigaHands offers detailed text annotations for all captured activities. With a total of 13k instructed motion sequences, 14k annotated motion clips, and 84k augmented text descriptions, it is larger than any other dataset. Furthermore, GigaHands includes 3.7 million bimanual 3D hand poses, represented by both 3D keypoints and MANO [86] hand meshes, comparable to the scale of Ego-Exo4D [33] – the current largest annotated 3D hand pose dataset. However, unlike Ego-Exo4D, which is annotated from selected, non-continuous frames, GigaHands provides continuous bimanual 3D hand poses and shapes. Each motion clip was captured using 51 camera views, resulting in 183M RGB frames (and 366M unique hand images). This multi-view setup enables new applications such as dynamic 3D radiance field reconstruction. Additionally,

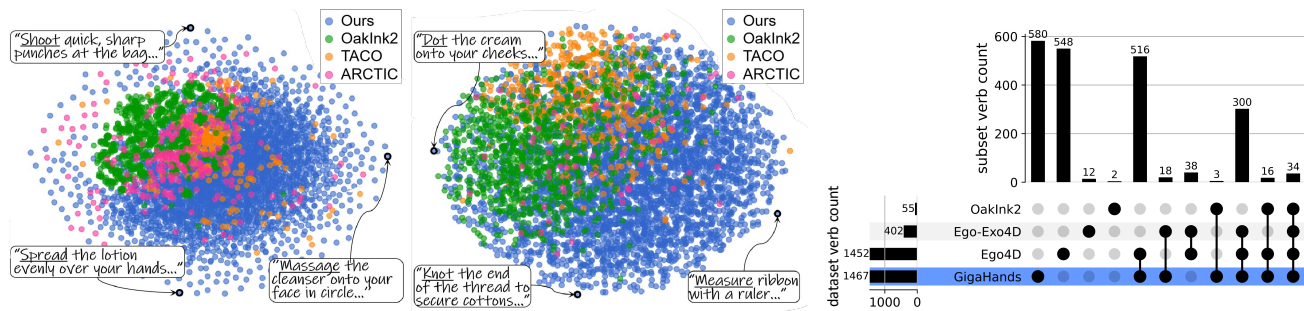


Figure 2. **Dataset Diversity.** The left and middle figures illustrate the diversity of pose and motion variations in GigaHands, visualized using t-SNE embeddings. Some points along the convex hull are highlighted with their corresponding text instructions, showcasing unique motions captured in our dataset. The right figure compares the verb sets among different datasets using an UpSet visualization [50]. Each column represents the number of verbs exclusive to specific subsets of datasets, indicated by the connected dots below the columns. The rows indicate the total verb count in each dataset. **GigaHands** contains more verbs and more exclusive verbs compared to other datasets.

the dataset contains annotations for hand/object segmentation masks, 3D object shape, pose and appearance, 2D/3D keypoints, and camera poses.

Diverse Hand Pose and Motion. Figure 2 (left) compares hand pose diversity across datasets using t-SNE [107] clustering of 3D keypoints of both hands. Similarly, Figure 2 (middle) shows hand motion diversity through t-SNE clustering of the latent codes of a Variational Autoencoder trained for motion reconstruction. GigaHands exhibits significantly greater diversity than existing datasets in both bimanual hand pose and motion. Additionally, datapoints near the boundary of our dataset seem to correspond to text annotations that are unique in our datasets, demonstrating that diverse verbs contribute to diverse poses and motions.

Diverse Text Annotation. GigaHands contains activities that are more diverse than existing datasets [32, 33, 116] as measured using verb counts in Figure 2 (right). Since each verb corresponds to an activity, a diverse set of verb indicates activity diversity. GigaHands contains the most verbs (1467) with 580 of them being unique to our dataset. The key to achieving this diversity is our *instruct-to-annotate* strategy (see Section 4.3) for automated labeled activity sourcing. Please refer to the supp. document for the source of verbs and detailed verb comparison.

Diverse Objects. Figure 3 shows some of the tabletop objects included in GigaHands. We provide 3D meshes [2] with associated textures for rigid objects and multi-view segmentation masks for non-rigid objects. Different from most datasets, we also include small-scale objects (e.g. pens) that are difficult to capture. We employ single-view reconstruction [3] to obtain meshes for these small objects. In total, the dataset includes 417 objects, with 310 multi-view scanned meshes and 31 single-view generated meshes.

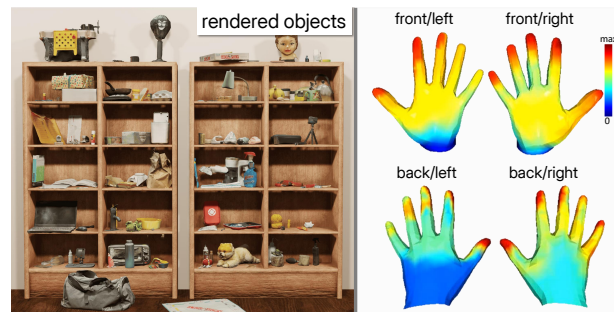


Figure 3. **Diverse Objects and Frequent Hand Contact Regions.** GigaHands provide objects (left) spanning diverse scenarios, including cooking, office working, crafting, entertainment, and housework. The diverse activities result in contact regions (right) spanning both the front and back of both hands.

Diverse Contacts. Since we have 3D meshes both hands and objects, we can use them to estimate **contact maps** on both hands, following the approach in [101]. Figure 3 shows the accumulated contact regions from randomly sampled frames across the dataset. The results reveal diverse contact areas, including regions between fingers and on the back of the hands (e.g. punching uses the backs of hands). The right hand interacts with objects more frequently than the left, since a majority of our subjects were right-handed.

4. Dataset Acquisition

We present our data acquisition pipeline, which balances realism, diversity, and accuracy while reducing annotation effort. Our procedural pipeline, called *Instruct-to-Annotate* (see Figure 4), consists of several key components:

- **Procedural instruction elicitation** (Section 4.1) for generating instructions to guide participants during data capture.
- **Filming** (Section 4.2) using a markerless multi-camera system to record high-quality hand activity sequences.

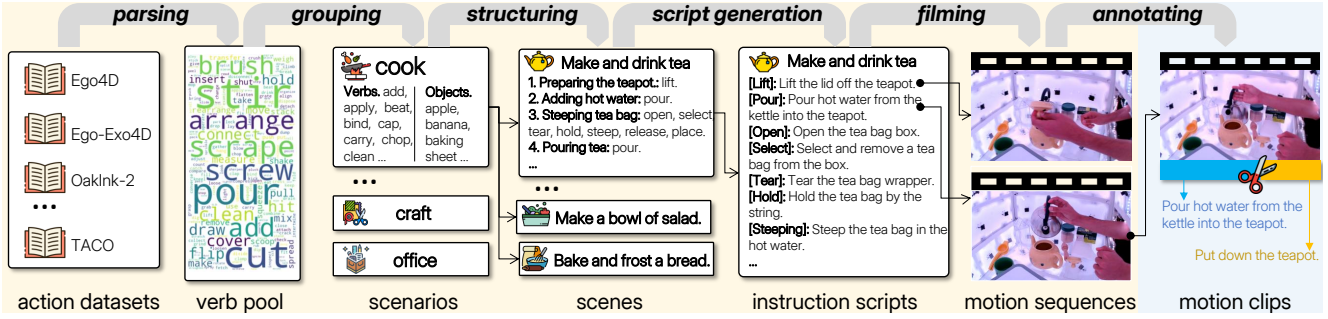


Figure 4. **Instruct-to-Annotate Pipeline.** The instruction elicitation process (left yellow block) creates atomic action-level instruction scripts in a temporally smooth order, structured within scenes. This is achieved by parsing action datasets, grouping verbs into a pool, structuring scenarios, and generating scene scripts. During filming, subjects act according to these scripts, producing recorded motion sequences. Annotators then process these sequences (right blue block) by segmenting them into clips and annotating unscripted motions.

- **Motion annotation and text augmentation** (Section 4.3) to refine sequences by breaking them into shorter clips with accurate annotations.
- **Hand motion** (Section 4.4) and **object motion estimation** (Section 4.5) for detailed 3D hand and object motions from the captured multi-view RGB videos.

4.1. Instruction Elicitation

To ensure diverse hand activities and reduce annotation effort, the *Instruct-to-Annotate* pipeline starts with a procedural instruction elicitation protocol. We began by sourcing atomic actions from multiple datasets [32, 33, 58, 116], extracting a pool of verbs corresponding to those actions. To elicit subjects to perform these verbs (actions), we manually associated each verb with multiple objects and, with the assistance of an LLM [4], grouped the verbs and objects into different scenarios such as cooking and eating, office work, crafting, entertainment, and housework. We then structured these scenarios into scenes where the objects could co-occur. Using LLM, we organized the scenes into lists of activities that utilized the objects and verbs in a temporally smooth order and automatically generated detailed instruction scripts. These scripts comprise 5 scenarios, 25 scenes, 191 activities, and 1370 instructions containing a total of 533 verbs. Please see the supp. document for details.

4.2. Filming

Hardware. To capture our dataset, we use a custom-designed multi-camera tabletop capture system that consists of 51 RGB cameras uniformly arranged within a cubic capture volume, with each face of the cube containing a 3×3 grid of cameras evenly illuminated by LED lights. Inside the cube, a transparent glass surface serves as a supportive platform for objects. Each camera records at 30fps with a resolution of 1280×720 . Cameras are software-synchronized, with the temporal phase misalignment being less than 3 ms. Camera intrinsics and extrinsics are obtained using COLMAP [87, 88] aided by fiducial markers.

Filming Process. During filming, subjects perform actions according to the provided instructions, keeping both hands within the filming area. Instructions are converted to audio and played sequentially to guide the participants, while an operator controls the capture by playing each audio instruction and recording each motion sequence. If the ending state of a performance does not align with the next instruction, a corrective instruction will be given and re-recorded to ensure smooth transitions and reduce annotation efforts. This approach ensures all recorded sequences correspond to a pre-scripted or recorded instruction.

4.3. Action Annotation & Augmentation

Though each filmed motion sequence was paired with an instruction, manual annotation was still necessary for two reasons. First, the instructions generated by the LLM occasionally contained inconsistencies or hallucinations. Second, participants might misunderstand the instructions or add unintentional actions during filming. To encourage subjects to act freely and naturally, we intentionally retained these recordings. Annotators then split these sequences into individual clips and annotated any actions not included in the original instructions. This process resulted in a more accurate dataset with matched (description, clip) pairs. In total, we refined the 13k motion sequences into 14k motion clips.

To further enhance the text descriptions, we used the LLM to rephrase each description 5 times, providing multiple textual variations for each motion clip. This expanded our 14k motion clips into 84k motion-text pairs containing 1,467 unique verbs. The augmentation prompts and examples are provided in the supp. document.

4.4. Hand Motion Estimation

GigaHands provides detailed hand motion data, including 2D and 3D hand keypoints and MANO [86] meshes for both hands. Since existing hand shape and pose estimation did not work well enough, we built our own hybrid

method. We begin by obtaining bounding boxes for the hands across videos using YOLOv8 [85]. Then, 2D keypoints are extracted from the MANO meshes estimated with HaMeR [77] and handedness (left or right) is determined with ViTPose [110]. HaMeR meshes cannot directly be used since they lack accurate depth. With camera parameters, 2D keypoints are triangulated across views [97] to obtain accurate 3D positions. To ensure temporal smoothness in the sequences, the one-euro filter [16] is applied to the both 2D and 3D keypoints. With the bounding box and 2D & 3D keypoints, we fit the MANO parametric hand model [86] following the EasyMoCap pipeline [1]. This results in a fully automated pipeline for extracting coherent and accurate hand motions. Details and the evaluation of each step are provided in the supp. document.

4.5. Object Motion Estimation

GigaHands also provides 3D object motions represented as 3D shape and 6D pose. Given target meshes obtained from pre-scanning or single-view reconstruction, we track these objects with multi-view constraints. First, we segment the target objects from the background in each view. We detect salient objects across the video using DINOv2 [74] on frames subsampled at 1 fps. Using both object text descriptions and rendered template meshes from multiple views, we select the top-k bounding boxes most aligned with the template mesh throughout the video with Grounding DINO [56]. To eliminate false positives, we use OpenCLIP [41] to filter out boxes aligned with negative prompts. With the positive bounding boxes as the prompt, we use SAM2 [84] to segment object masks throughout the video.

Since no existing method for object pose estimation worked well [75], we decided to build our own robust method that exploits our dense multi-view setting. We use a differentiable rendering approach [83] supervised by the multi-view masks. To initialize the translation and rotation, we first build a radiance field using Instant-NGP [68]. We initialize the object’s translation using the center of the density field. To address potential object symmetries, we use object appearance to initialize the rotation. Following FoundPose [75], we render the template mesh with multiple initial rotations and use DINOv2 features to find the best match, aggregating cross-view information. This process yields precise object motion estimates even in complex, cluttered scenes. More details and the evaluation of each step are provided in the supp. document.

5. Applications & Experiments

We demonstrate the utility of GigaHands in several applications that require large-scale data, including text-driven hand motion synthesis (Section 5.1), motion captioning for our dataset (Section 5.2) and for in-the-wild datasets (Section 5.2), and dynamic radiance field reconstruction (Sec-

Table 2. Quantitative results for text-driven motion synthesis with models trained on different datasets. *upper bound* indicates performance calculated with the ground truth. We report the mean of 20 evaluations, and $\rightarrow$ means the closer to the upper bound the better. The model trained on GigaHands performs best on most metrics.

Dataset	R Precision(%) $\uparrow$			MM Dist. $\downarrow$	FID $\downarrow$	Div. $\rightarrow$	MM. $\uparrow$
	@1	@2	@3				
upper bound	64.4	86.2	89.4	2.86	0.045	14.2	-
TACO [58]	18.9	37.7	52.9	7.39	11.0	11.1	6.83
upper bound	50.4	71.2	81.1	3.67	0.022	9.30	-
OakInk2 [116]	17.9	31.7	47.9	7.75	19.6	6.88	3.45
upper bound	77.4	88.8	91.3	2.96	0.002	11.9	-
GigaHands	31.2	44.7	53.1	6.68	4.70	10.5	9.11

tion 5.3). More results can be found in the supp. document.

5.1. Text-driven Hand Motion Synthesis

Generating diverse and complex motions is crucial for training virtual agents and robotic manipulation. We use GigaHands to train models for text-driven hand motion synthesis, overcoming the limitations of approaches that rely on strict conditions [116] or generate only simple skills [17] due to constrained data. We demonstrate that the scale of GigaHands leads to better performance compared to other datasets [58, 116] and smaller subsets of GigaHands.

Training and Evaluation Protocol. We split GigaHands into train, test, and val sets with a ratio of 16:3:1. From among our annotations, we chose 42 3D keypoints for both hands as the representation for model training (see the supp. document for an analysis of alternative representations). To evaluate the naturalness, diversity, and alignment of generated motions with textual descriptions, we use metrics from [34], including R-Precision, Multimodal Distance (MM Dist), Fréchet Inception Distance (FID), Diversity (Div.), and Multimodality (MM.). For computing FID and Div., we train a motion autoencoder to define a compressed motion space. For R-Precision, MM Dist, and MM., we employ contrastive learning to create a joint motion-text embedding space, following [34]. Feature extractors are trained independently for each dataset and subset.

Results. We report text to hand motion synthesis performance using the T2M-GPT [119] backbone in Table 2 (see the supp. document for more results on GRU-based [35] and diffusion [103] architectures). Since the evaluation embedding spaces are trained on different test sets, we also evaluate the metrics on ground truth test data to indicate upper-bound performance for each test set. Models trained on GigaHands outperform others on all metrics except MM Dist., with higher R-Precision due to better text-motion alignment. Compared to OakInk2 (which also includes textual descriptions for hand motions), GigaHands achieves significantly better FID, Diversity, and Multimodality scores,



Figure 5. Generated motions from models trained on different datasets. Texts highlighted in green, orange, and blue come from the OakInk2, TACO, and GigaHands datasets. In the bottom two rows, hand meshes highlighted in these colors are generated by models trained on the corresponding datasets. The model trained on GigaHands can generate diverse motions from a single text (right four columns) and accurate motion with text from other datasets (left two columns). Darker color indicates later frame in the sequence.

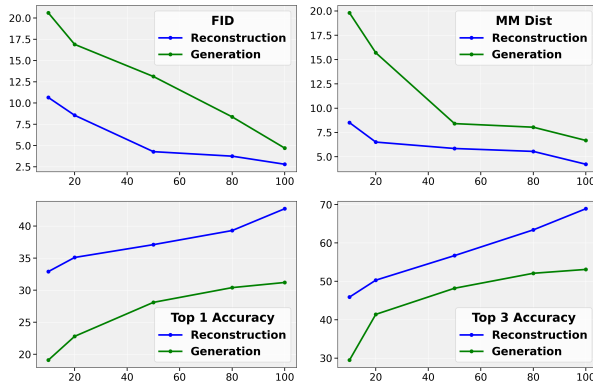


Figure 6. Effect of dataset size on motion reconstruction and text-to-motion generation performance. The x-axis shows the percentage of training data used (10%, 20%, 50%, 80%, and 100%), and the y-axis displays performance metrics: FID, MM Dist., Top-1, and Top-3 accuracy. Larger datasets consistently improve performance across all metrics, highlighting the benefits of increased data scale.

likely due to its greater dataset diversity. Figure 5 shows hand motions generated by T2M-GPT trained with GigaHands using text inputs from the test sets of GigaHands, OakInk and TACO. Even without object geometry input, our model generates reasonable hand shapes and poses for object manipulation, benefiting from the diversity of objects in our dataset. Our model can also generate reasonable motion using text from other datasets, showing the comprehensiveness of GigaHands.

Effect of Data Scale. Figure 6 illustrates the impact of dataset size on motion reconstruction and text-to-motion generation. Based on the T2M-GPT architecture, we train a motion VQ-VAE for reconstruction and a generative pre-trained transformer model for text-to-motion generation using 10%, 20%, 50%, 80%, and 100% of the training set, evaluating on the same test set. We evaluate FID, MM Dist., and Top 1 & Top 3 accuracies. Most metrics continually improve with larger datasets, demonstrating the value

of larger-scale data.

5.2. Hand Motion Captioning

Captioning human hand motion helps interpret action intent. We train a motion captioning model on GigaHands and generate captions for its test set motions; we also show examples of captioning in-the-wild datasets.

Training and Evaluation Protocol. We adopt TM2T [35] as the model backbone. For text-motion alignment evaluation, we utilize R-Precision and MM Dist., employing the same embedding space described in Section 5.1. To evaluate alignment with the ground-truth text annotations, we use BLEU [76], ROUGE [53], and BERTScore [121]. Since our dataset has diverse verbs, we evaluate text diversity with distinct-n [51] and Pairwise BLEU [94]. Since TACO lacks text annotations, we use its triplet labels (`<action, tool, object>`) as scripts, transforming motion captioning in TACO into a classification task.

Table 3. Quantitative evaluation for motion captioning with models trained on different datasets. GigaHands performs best on most metrics.

Datasets	RP@1(%)↑	RP@2(%)↑	RP@3(%)↑	MM Dist↓
upper bound	64.4	86.2	94.2	2.86
TACO [58]	43.5	59.6	68.5	5.84
upper bound	51.1	69.5	79.6	3.67
Oakink2 [116]	40.4	58.8	68.2	4.55
upper bound	75.3	89.1	93.9	2.87
GigaHands	57.0	66.1	69.8	5.37

Table 4. Pairwise-BLEU, BLEU@4, ROUGE, distinct-n, and BERTScore for motion captioning with models trained on different datasets. GigaHands performs best on most metrics.

Datasets	P-B↓	B@4↑	R↑	dist.-1(%)↑	dist.-2(%)↑	BScore↑
TACO [58]	4.59	39.4	61.2	2.03	6.35	57.1
Oakink2 [116]	8.21	39.9	56.3	2.76	6.82	35.3
GigaHands	0.916	43.1	57.7	15.3	36.9	55.4

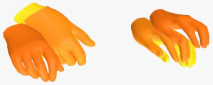
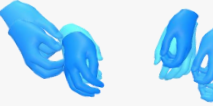
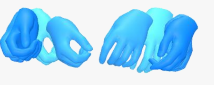
	motion									
										
	GT Stir in the yellow mug with the spoon.		(Cut, Knife, Plate)		Pour hot water from the kettle into the teapot		Uncap the pen		Grab the watering bottle	
#1	Stir in the blue cup with the rod		(Cut, Knife, Plate)		Fill the teacup with tea from the kettle		Twist off the pen		Take hold of the watering bottle	
#2	Stir the tea and sugar together with the teaspoon until well mixed		Chop the apple into smaller segments		Dispense the tea into the teacup from the kettle		Remove the cap from the highlighter		Pick up the watering bottle	

Figure 7. **Motion captioning results with different datasets.** Each column shows a motion sequence, its ground truth text description, and two generated texts. Hand motions highlighted in green, orange, and blue come from OakInk2, TACO, and GigaHands, respectively. Texts highlighted in these colors are generated by models trained on the corresponding datasets. The model trained on GigaHands generates diverse captions from a single motion (right three columns) and accurately captions motions from other datasets (left two columns).

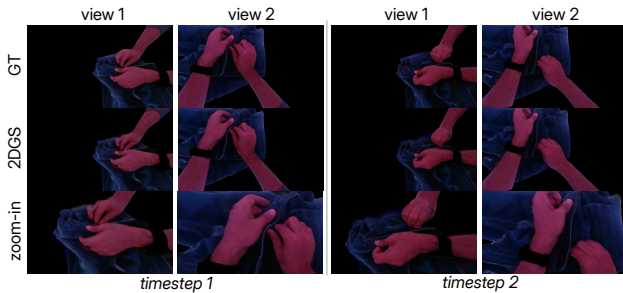


Figure 8. Synthesized test views using 2DGS for the motion ‘zip up the pants,’ displayed at two timesteps and from two viewpoints. The synthesized views faithfully reconstruct the scene by leveraging the ample camera views provided by GigaHands.

Hand Motion Captioning for GigaHands Figure 7 shows examples of generated 3D motion captions. The model trained on GigaHands generates diverse and accurate captions. Though objects are not present in the input, the model still generates captions which include reasonable object descriptions. Table 3 and Table 4 report the quantitative performance of models trained on different datasets [58, 116]. The model trained on GigaHands achieves superior motion-text alignment, particularly in Pairwise BLEU and distinct-n, which reflect caption diversity. While our accuracy is marginally lower compared to the simple triplet captioning in TACO, our approach produces richer and more varied captions.

3D Hand Motion Captioning for Unlabeled Datasets GigaHands’s large variety motions and associated annotations allows models trained on it to caption other 3D hand motion datasets. Fig. 7 shows some examples using the model trained on GigaHands to caption motions from TACO and OakInk2 datasets after aligning their motion range with GigaHands. This model trained with GigaHands generates finer-granularity captions for these other datasets.

5.3. Dynamic Radiance Field Reconstruction

Unlike existing hand activity datasets, GigaHands provides dense camera views. This enables new applications, such as

building dynamic radiance fields. From the 51 views provided, we remove 12 camera views due to lighting issues and randomly select one view as the test view, resulting in 38 views used for training. We segment consistent object and hand masks throughout the video as described in Section 4.5. We then fit 2DGS [40] for frame-wise radiance field reconstruction, initializing each frame with the previous frame for temporal consistency. Figure 8 shows an example of synthesized test views for a motion clip. Notably, the pants are a non-rigid object that cannot be easily tracked, but GigaHands provides ample camera views for capturing non-rigid objects (more results in supp. document).

6. Conclusion

We present GigaHands, a massive annotated dataset of bi-manual hand activities. The dataset contains 14k motion clips with 3D hand and object motions from 56 subjects interacting with 417 real-world objects, all captured from 51 camera views. It provides 183 million unique image frames and 84k textual descriptions, enabling a wide range of applications. We demonstrated how the scale and diversity of the dataset benefit text-driven motion synthesis and motion captioning, both within the dataset and on other data. Additionally, we have shown that GigaHands enables dynamic radiance field reconstruction, opening possibilities for downstream tasks.

Limitations and Future Directions. Despite its scale and diversity, GigaHands has limitations. The studio setting confines data collection to a limited space, making it challenging to accurately capture motions that require larger environments. While we can track rigid objects and object parts, fully automatic tracking of articulated and non-rigid objects remains challenging. Moreover, although we have showcased applications in motion synthesis and captioning, further research could explore how the dataset’s scale and diversity can enhance robotic manipulation and human-computer interaction tasks.

Acknowledgement

This research was supported by AFOSR grant FA9550-21-1-0214, NSF CAREER grant #2143576, and ONR DURIP grant N00014-23-1-2804. We would like to thank the OpenAI Research Access Program for API support and extend our gratitude to Ellie Pavlick, Tianran Zhang, Carmen Yu, Angela Xing, Chandradeep Pokhariya, Sudarshan Harithas, Hongyu Li, Chaerin Min, Xindi Qu, Xiaoquan Liu, Hao Sun, Melvin He and Brandon Woodard.

References

- [1] EasyMocap - make human motion capture easier. Github, 2021. 6
- [2] AR Code, Augmented Reality QR Codes — ar-code.com. <https://ar-code.com/>, 2024. [Accessed 05-11-2024]. 4
- [3] Meshy Docs — docs.meshy.ai. <https://docs.meshy.ai/image-to-3d>, 2024. [Accessed 13-11-2024]. 4
- [4] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 5
- [5] Heni Ben Amor, Oliver Kroemer, Ulrich Hillenbrand, Gerhard Neumann, and Jan Peters. Generalization of human grasping for multi-fingered robot hands. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2043–2050. IEEE, 2012. 2
- [6] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 3
- [7] Yunfei Bai and C Karen Liu. Dexterous manipulation using both palm and fingers. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1560–1565. IEEE, 2014. 2
- [8] 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, 3
- [9] Samarth Brahmbhatt, Cusuh Ham, Charles C Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8709–8719, 2019. 3
- [10] Samarth Brahmbhatt, Ankur Handa, James Hays, and Dieter Fox. Contactgrasp: Functional multi-finger grasp synthesis from contact. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2386–2393. IEEE, 2019. 2
- [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. 2
- [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. 2
- [14] Gereon H Büscher, Risto Kõiva, Carsten Schürmann, Robert Haschke, and Helge J Ritter. Flexible and stretchable fabric-based tactile sensor. *Robotics and Autonomous Systems*, 63:244–252, 2015. 3
- [15] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12417–12426, 2021. 3
- [16] Géry Casiez, Nicolas Roussel, and Daniel Vogel. 1€ filter: a simple speed-based low-pass filter for noisy input in interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2527–2530, 2012. 6
- [17] Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2hoi: Text-guided 3d motion generation for hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1577–1585, 2024. 3, 6
- [18] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9044–9053, 2021. 3
- [19] Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20577–20586, 2022. 2
- [20] Sammy Christen, Shreyas Hampali, Fadime Sener, Edoardo Remelli, Tomas Hodan, Eric Sauser, Shugao Ma, and Bugra Tekin. Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions. *arXiv preprint arXiv:2403.17827*, 2024. 2
- [21] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5031–5041, 2020. 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, 3
- [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, 3
- [24] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. *Advances in Neural Information Processing Systems*, 35:13745–13758, 2022. 3
- [25] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36, 2024. 2
- [26] Christian RG Dreher, Mirko Wächter, and Tamim Asfour. Learning object-action relations from bimanual human demonstration using graph networks. *IEEE Robotics and Automation Letters*, 5(1):187–194, 2019. 3
- [27] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2
- [28] 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. 2, 3
- [29] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 409–419, 2018. 2, 3
- [30] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The” something something” video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 2, 3
- [31] Patrick Grady, Chengcheng Tang, Samarth Brahmabhatt, Christopher D Twigg, Chengde Wan, James Hays, and Charles C Kemp. Pressurevision: estimating hand pressure from a single rgb image. In *European Conference on Computer Vision*, pages 328–345. Springer, 2022. 3
- [32] 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, 3, 4, 5
- [33] 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, 4, 5
- [34] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5152–5161, 2022. 6
- [35] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *ECCV*, 2022. 6, 7
- [36] 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
- [37] 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
- [38] Shangchen Han, Po-chen Wu, Yubo Zhang, Beibei Liu, Linguang Zhang, Zheng Wang, Weiguang Si, Peizhao Zhang, Yujun Cai, Tomas Hodan, et al. Umetrack: Unified multi-view end-to-end hand tracking for vr. In *SIGGRAPH Asia 2022 conference papers*, pages 1–9, 2022. 2
- [39] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11807–11816, 2019. 2, 3
- [40] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 8
- [41] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hananeh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 6
- [42] 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
- [43] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, Au-*

- gust 23–28, 2020, *Proceedings, Part IX 16*, pages 196–214. Springer, 2020. 3
- [44] Yitong Jin, Zhiping Qiu, Yi Shi, Shuangpeng Sun, Chongwu Wang, Donghao Pan, Jiachen Zhao, Zhenghao Liang, Yuan Wang, Xiaobing Li, et al. Audio matters too! enhancing markerless motion capture with audio signals for string performance capture. *ACM Transactions on Graphics (TOG)*, 43(4):1–10, 2024. 3
- [45] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8320–8329, 2018. 3
- [46] 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. 2
- [47] Franziska Krebs, Andre Meixner, Isabel Patzer, and Tamim Asfour. The kit bimanual manipulation dataset. In *2020 IEEE-RAS 20th International Conference on Humanoid Robots (Humanoids)*, pages 499–506. IEEE, 2021. 3
- [48] 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
- [49] Gilwoo Lee, Zhiwei Deng, Shugao Ma, Takaaki Shiratori, Siddhartha S Srinivasa, and Yaser Sheikh. Talking with hands 16.2 m: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 763–772, 2019. 3
- [50] Alexander Lex, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, and Hanspeter Pfister. Upset: visualization of intersecting sets. *IEEE transactions on visualization and computer graphics*, 20(12):1983–1992, 2014. 4
- [51] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*, 2015. 7
- [52] Yin Li, Miao Liu, and James M Rehg. In the eye of the beholder: Gaze and actions in first person video. *IEEE transactions on pattern analysis and machine intelligence*, 45(6):6731–6747, 2021. 2, 3
- [53] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 7
- [54] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [55] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European conference on computer vision*, pages 612–630. Springer, 2022. 2, 3
- [56] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 6
- [57] 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. 2, 3
- [58] 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. 2, 3, 5, 6, 7, 8
- [59] 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
- [60] Priyanka Mandikal and Kristen Grauman. Learning dexterous grasping with object-centric visual affordances. In *2021 IEEE international conference on robotics and automation (ICRA)*, pages 6169–6176. IEEE, 2021. 2
- [61] R McKee, D McKee, D Alexander, and E Paillat. Nz sign language exercises. deaf studies department of victoria university of wellington, 2024. 3
- [62] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019. 2, 3
- [63] Andrew T Miller and Peter K Allen. Graspit! a versatile simulator for robotic grasping. *IEEE Robotics & Automation Magazine*, 11(4):110–122, 2004. 2
- [64] Chaerin Min and Srinath Sridhar. Genheld: Generating and editing handheld objects. *arXiv preprint arXiv:2406.05059*, 2024. 3
- [65] Gyeongsik Moon, Shou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 548–564. Springer, 2020. 3
- [66] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *Proceedings of the IEEE international conference on computer vision*, pages 1154–1163, 2017. 2, 3
- [67] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and

- Christian Theobalt. Generated hands for real-time 3d hand tracking from monocular rgb. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 49–59, 2018. 3
- [68] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. 6
- [69] 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
- [70] Supreeth Narasimhaswamy, Zhengwei Wei, Yang Wang, Justin Zhang, and Minh Hoai. Contextual attention for hand detection in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9567–9576, 2019. 3
- [71] Evonne Ng, Javier Romero, Timur Bagautdinov, Shaojie Bai, Trevor Darrell, Angjoo Kanazawa, and Alexander Richard. From audio to photoreal embodiment: Synthesizing humans in conversations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1001–1010, 2024. 3
- [72] Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. Assemblyhands: Towards egocentric activity understanding via 3d hand pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12999–13008, 2023. 3
- [73] 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. 2
- [74] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 6
- [75] Evin Pınar Örnek, Yann Labbé, Bugra Tekin, Lingni Ma, Cem Keskin, Christian Forster, and Tomas Hodan. Foundpose: Unseen object pose estimation with foundation features. *arXiv preprint arXiv:2311.18809*, 2023. 6
- [76] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 7
- [77] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2024. 3, 6
- [78] Tu-Hoa Pham, Nikolaos Kyriazis, Antonis A Argyros, and Abderrahmane Kheddar. Hand-object contact force estimation from markerless visual tracking. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2883–2896, 2017. 3
- [79] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2847–2854. IEEE, 2012. 2, 3
- [80] Chandradeep Pokhariya, Ishaan Nikhil Shah, Angela Xing, Zekun Li, Kefan Chen, Avinash Sharma, and Srinath Sridhar. Manus: Markerless grasp capture using articulated 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2197–2208, 2024. 3
- [81] 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. 2
- [82] Chen Qian, Xiao Sun, Yichen Wei, Xiaou Tang, and Jian Sun. Realtime and robust hand tracking from depth. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1106–1113, 2014. 3
- [83] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 6
- [84] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 6
- [85] Dillon Reis, Jordan Kupec, Jacqueline Hong, and Ahmad Daoudi. Real-time flying object detection with yolov8. *arXiv preprint arXiv:2305.09972*, 2023. 6
- [86] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022. 2, 3, 5, 6
- [87] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5
- [88] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 5
- [89] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2
- [90] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022. 3

- [91] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020. 2, 3
- [92] Toby Sharp, Cem Keskin, Duncan Robertson, Jonathan Taylor, Jamie Shotton, David Kim, Christoph Rhemann, Ido Leichter, Alon Vinnikov, Yichen Wei, et al. Accurate, robust, and flexible real-time hand tracking. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 3633–3642, 2015. 3
- [93] Qijin She, Ruizhen Hu, Juzhan Xu, Min Liu, Kai Xu, and Hui Huang. Learning high-dof reaching-and-grasping via dynamic representation of gripper-object interaction. *arXiv preprint arXiv:2204.13998*, 2022. 2
- [94] Tianxiao Shen, Myle Ott, Michael Auli, and Marc’Aurelio Ranzato. Mixture models for diverse machine translation: Tricks of the trade. In *International conference on machine learning*, pages 5719–5728. PMLR, 2019. 7
- [95] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018. 2
- [96] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1145–1153, 2017. 3
- [97] Srinath Sridhar, Antti Oulasvirta, and Christian Theobalt. Interactive markerless articulated hand motion tracking using rgb and depth data. In *Proceedings of the IEEE international conference on computer vision*, pages 2456–2463, 2013. 3, 6
- [98] Srinath Sridhar, Franziska Mueller, Michael Zollhöfer, Dan Casas, Antti Oulasvirta, and Christian Theobalt. Real-time joint tracking of a hand manipulating an object from rgb-d input. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*, pages 294–310. Springer, 2016. 3
- [99] Xiao Sun, Yichen Wei, Shuang Liang, Xiaou Tang, and Jian Sun. Cascaded hand pose regression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 824–832, 2015. 3
- [100] Subramanian Sundaram, Petr Kellnhofer, Yunzhu Li, Jun-Yan Zhu, Antonio Torralba, and Wojciech Matusik. Learning the signatures of the human grasp using a scalable tactile glove. *Nature*, 569(7758):698–702, 2019. 3
- [101] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. 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, Part IV 16*, pages 581–600. Springer, 2020. 2, 3, 4
- [102] Danhang Tang, Hyung Jin Chang, Alykhan Tejani, and Tae-Kyun Kim. Latent regression forest: Structured estimation of 3d articulated hand posture. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3786–3793, 2014. 3
- [103] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 6
- [104] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics (ToG)*, 33(5):1–10, 2014. 3
- [105] Dylan Turpin, Liquan Wang, Eric Heiden, Yun-Chun Chen, Miles Macklin, Stavros Tsogkas, Sven Dickinson, and Animesh Garg. Grasp’d: Differentiable contact-rich grasp synthesis for multi-fingered hands. In *European Conference on Computer Vision*, pages 201–221. Springer, 2022. 2
- [106] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision*, 118: 172–193, 2016. 3
- [107] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 4
- [108] Frank R Wilson. *The hand: How its use shapes the brain, language, and human culture*. Vintage, 1999. 2
- [109] Wei Xie, Zhipeng Yu, Zimeng Zhao, Binghui Zuo, and Yangang Wang. Hmdo: Markerless multi-view hand manipulation capture with deformable objects. *Graphical Models*, 127:101178, 2023. 3
- [110] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022. 6
- [111] Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. Oakink: A large-scale knowledge repository for understanding hand-object interaction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20953–20962, 2022. 2, 3
- [112] Ruolin Ye, Wenqiang Xu, Zhendong Xue, Tutian Tang, Yanfeng Wang, and Cewu Lu. H2o: A benchmark for visual human-human object handover analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15762–15771, 2021. 3
- [113] Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. The genea challenge 2022: A large evaluation of data-driven co-speech gesture generation. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 736–747, 2022. 3
- [114] Zhengdi Yu, Shaoli Huang, Yongkang Cheng, and Tolga Birdal. Signavatars: A large-scale 3d sign language holistic motion dataset and benchmark, 2024. 3
- [115] Shanxin Yuan, Qi Ye, Bjorn Stenger, Siddhant Jain, and Tae-Kyun Kim. Bighand2. 2m benchmark: Hand pose dataset and state of the art analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4866–4874, 2017. 3

- [116] Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 445–456, 2024. 2, 3, 4, 5, 6, 7, 8
- [117] Hui Zhang, Sammy Christen, Zicong Fan, Otmar Hilliges, and Jie Song. Grasppl: Generating grasping motions for diverse objects at scale. *arXiv preprint arXiv:2403.19649*, 2024. 2
- [118] Hui Zhang, Sammy Christen, Zicong Fan, Luocheng Zheng, Jemin Hwangbo, Jie Song, and Otmar Hilliges. Artigrasp: Physically plausible synthesis of bi-manual dexterous grasping and articulation. In *2024 International Conference on 3D Vision (3DV)*, pages 235–246. IEEE, 2024. 2
- [119] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14730–14740, 2023. 6
- [120] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *European conference on computer vision*, pages 180–200. Springer, 2022. 3
- [121] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. 7
- [122] Wenqian Zhang, Molin Huang, Yuxuan Zhou, Juze Zhang, Jingyi Yu, Jingya Wang, and Lan Xu. Both2hands: Inferring 3d hands from both text prompts and body dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2393–2404, 2024. 3
- [123] Zhongqun Zhang, Hengfei Wang, Ziwei Yu, Yihua Cheng, Angela Yao, and Hyung Jin Chang. Nl2contact: Natural language guided 3d hand-object contact modeling with diffusion model. *arXiv preprint arXiv:2407.12727*, 2024. 3
- [124] Chengfeng Zhao, Juze Zhang, Jiashen Du, Ziwei Shan, Junye Wang, Jingyi Yu, Jingya Wang, and Lan Xu. I’m hoi: Inertia-aware monocular capture of 3d human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 729–741, 2024. 3
- [125] Zehao Zhu, Jiashun Wang, Yuzhe Qin, Deqing Sun, Varun Jampani, and Xiaolong Wang. Contactart: Learning 3d interaction priors for category-level articulated object and hand poses estimation. In *2024 International Conference on 3D Vision (3DV)*, pages 201–212. IEEE, 2024. 3
- [126] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE international conference on computer vision*, pages 4903–4911, 2017. 2, 3
- [127] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 813–822, 2019. 3