OAKINK2 : A Dataset of Bimanual Hands-Object Manipulation in Complex Task Completion

Xinyu Zhan[1⋆] Lixin Yang[1⋆] Yifei Zhao[1] Kangrui Mao[1] Hanlin Xu[1]
Zenan Lin[2,‡] Kailin Li[1] Cewu Lu[1†]
[1]Shanghai Jiao Tong University, [2]South China University of Technology

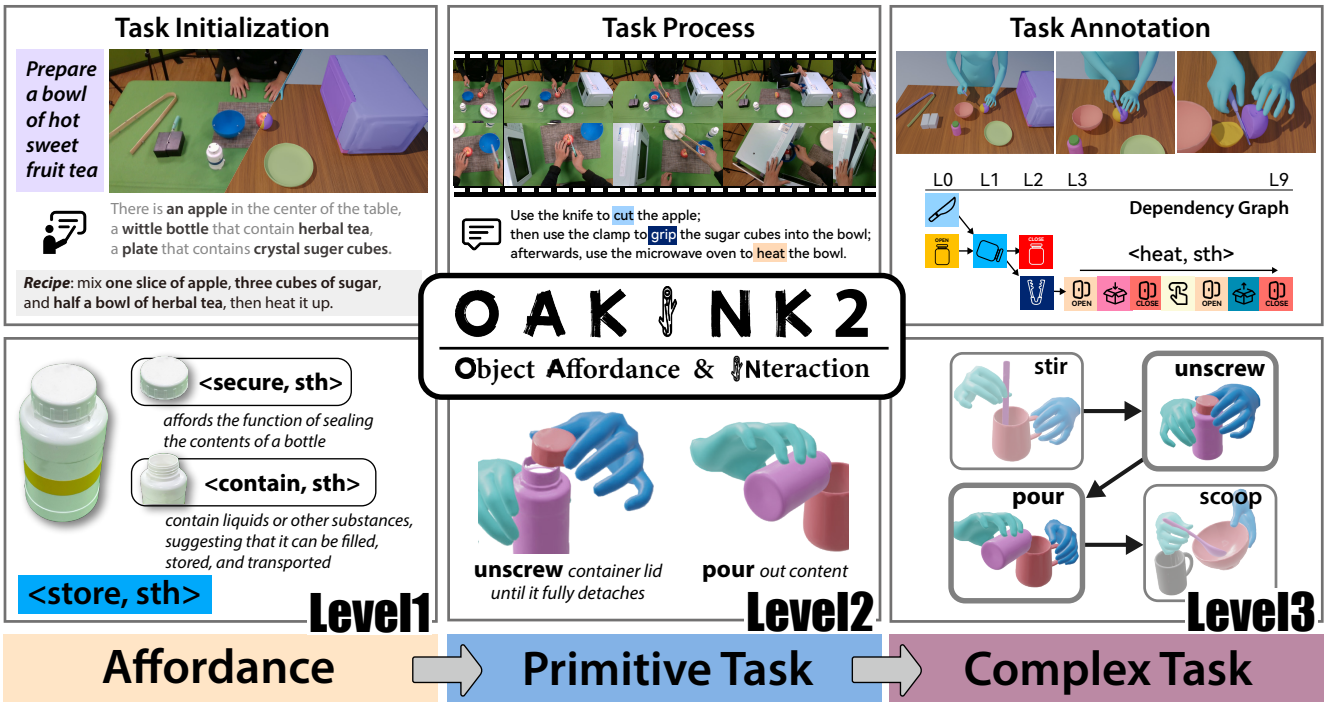

Figure 1. **An overview of the data and content of our proposed OAKINK2 dataset.** OAKINK2 dataset focuses on bimanual object manipulation tasks for complex daily activities. **1) The top row shows the data collection process**, including the task setup (top-left panel), human demonstration (top-center), and annotation (top-right). **2) The second row shows the three levels of abstraction constructed by OAKINK2 for complex tasks**, including the Affordance, Primitive Task, and Complex Task. OAKINK2 dataset provides allocentric and egocentric videos of human manipulation process, as well as the corresponding 3D-pose annotation and task specification.

Abstract

We present **OAKINK2**, *a dataset of bimanual object manipulation tasks for complex daily activities. In pursuit of constructing the complex tasks into a structured representation,* OAKINK2 *introduces three level of abstraction to organize the manipulation tasks:* ***Affordance***, ***Primitive Task***, *and* ***Complex Task***. OAKINK2 *features on an object-centric perspective for decoding the complex tasks, treating them as a sequence of object affordance fulfillment. The first level, Affordance, outlines the functionalities that objects in the scene can afford, the second level, Primitive Task, describes the minimal interaction units that humans interact with the object to achieve its affordance, and the third level, Complex Task, illustrates how Primitive Tasks are composed and interdependent.* OAKINK2 *dataset provides multi-view image streams and precise pose annotations for the human body, hands and various interacting objects. This extensive collection supports applications such as interaction reconstruction and motion synthe-*

⋆The first two authors contributed equally.
‡This work is done when Lin is an intern at SJTU.
†Cewu Lu is the corresponding author. He is the member of Qing Yuan Research Institute and MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, China.

sis. Based on the 3-level abstraction of OAKINK2*, we explore a task-oriented framework for Complex Task Completion (CTC). CTC aims to generate a sequence of bimanual manipulation to achieve task objectives. Within the CTC framework, we employ Large Language Models (LLMs) to decompose the complex task objectives into sequences of Primitive Tasks and have developed a Motion Fulfillment Model that generates bimanual hand motion for each Primitive Task.* OAKINK2 *datasets and models are available at* https://oakink.net/v2.

1. Introduction

Learning how humans achieve specific task objectives through diverse object manipulation behaviors has been a long-standing challenge. Recent data-driven approaches have made significant progress on this topic, including hand-object pose estimation [1, 7, 13, 20, 22–24, 36, 59], interaction synthesis [12, 15, 29, 51, 56, 63], and action imitation [47, 48]. However, the gap still exists for current methods to achieve a human-level understanding on object manipulation for complex task completion. In particular, humans possess a remarkable capacity to interact with specific objects in an appropriate sequence to achieve desired outcomes [30]. This inspires us to focus on the decomposition of hands-object interaction in complex manipulation tasks into sequential units.

Tracing prior research, the advancement in hand-object interaction understanding is inseparable from the emergence of a series of hand-object interaction datasets [3, 8, 12, 14, 19, 22, 28, 31, 37, 42, 48, 50, 60, 64] to support data-driven methods. A noteworthy example among these datasets is OakInk [60]. OakInk analyzed object affordances (*i.e.* functional properties of objects/object-parts [16]) and collected *human-centric* grasping interaction driven by intents to utilize these affordances. The term: *Oak* is for object affordance knowledge, and *Ink* for interaction knowledge. Nevertheless, the previous OakInk has two major limitations: 1) it lacks human demonstrations that cover the process of fulfilling those affordances, and 2) it lacks complex manipulation tasks that involve multiple object affordances.

In this paper, we present OAKINK2, extending the data and methodology of the previous OakInk. In order to manage the inherent complexity in complex manipulation tasks, OAKINK2 adopts an *object-centric* perspective and constructs three levels of abstraction upon manipulation tasks:

1) ***Affordance***: object/object-part level functionalities that enable manipulation. For example, a bottle cap affords securing and unsecuring of the content in the bottle.

2) ***Primitive Task (Primitive)***: a "minimal" sequence of hand-object interaction that fulfills a given object's affordance. For instance, to fulfill the affordance: securing, one needs to either `screw` or `press` the cap onto the bottle's opening to form a seal that prevents leaking.

3) ***Complex Task***: sequential combination of *Primitives* to address the long-horizon and multi-goals manipulation tasks. Tasks are characterized as "complex" for their goal requires more than one object affordance. *Complex Tasks* also detail the **dependencies** among the *Primitives* and dictate the **order** in which they are executed. To illustrate, to pour the fluid from a sealed bottle, one must first `unscrew` the cap and then `pour` out the liquid.

In this way, OAKINK2 delineates *Complex Tasks* as directed acyclic graphs, hereafter referred to as ***Primitive* Dependency Graphs (PDG)**. Within these graphs, each node represents a *Primitive*, serving to fulfill a specific affordance. The directed edges illustrate the sequence in which *Primitives* must be executed to achieve task completion.

Build upon the above methodology, OAKINK2 introduces a large-scale dataset for bimanual object manipulation. It encompasses human demonstrations for complex task completion, with multi-view image streams and paired pose annotations for human body, hands and objects. OAKINK2 contains 627 sequences of real-world bimanual manipulation sequences, where 264 of these sequences are for *Complex Tasks*. These sequences contain 4.01M frames from four different views (one egocentric and three allocentric views). The dataset includes four manipulation scenarios, 75 objects and 9 invited subjects in total.

The versatile and task-driven nature of OAKINK2 enables a wide range of applications. In this paper, we focus on the task and motion planning for Complex Task Completion (CTC). CTC involves two notable components: **1)** text-based *Complex Task* decomposition using *Primitives* and **2)** task-aware motion generation to fulfill each *Primitive*. For the first component, we design a task interpreter with Large Language Model (LLM) that can generate the PDG and program the execution order of these *Primitives*, based on textual descriptions of the *Complex Tasks*. For the latter component, we propose a generalist Task-aware Motion Fulfillment model (TaMF) to generate the hand motion at *Primitive* level, based on the task-related object trajectory.

In summary, our contributions are as follows:

- We build an object-centric, three-level abstraction to structure and understand complex manipulation tasks, *i.e.* *Affordance*, *Primitive* to fulfill affordance, and *Complex Task* with *Primitive* dependencies.
- We introduce OAKINK2, a large-scale real-world dataset for bimanual object manipulation with human demonstrations for both *Primitives* and *Complex Tasks*.
- We propose a task-oriented framework, CTC, for complex task and motion planning. CTC consists of a LLM-based task interpreter for *Complex Task* decomposition and a diffusion-based motion generator for *Primitive* fulfillment.

2. Related Works

Hand-Object Interaction Datasets. The recent research community has witnessed the emergence of numerous datasets on hand-object interactions. Earlier datasets [3, 12, 22] focused on static hand-object interactions with limited diversity. More recent datasets [8, 14, 19, 31, 38, 50, 64] captured dynamic hand-object interactions, covering bimanual interactions [14, 31] and interactions with articulated bodies [14, 64]. We pay particular attention to interaction datasets related to object affordances. [12] expressed affordances in grasp type labels. [3, 14, 50] collected intention labels for interactions. [28, 60] studied object affordance-based hand-object interaction and collected object segmentations and affordance labels. [38] studied hand-object interactions in tool-action-object pairs. Our proposed OAKINK2 captures both human demonstrations for minimal interaction fulfilling object affordance as *Primitive*, and demonstrations for *Complex Task* where these affordances are fulfilled in specific order constrained by their dependencies.

Decomposition of Manipulation Tasks. Decomposing complex manipulation tasks into multiple building blocks across different hierarchies represents a widely adopted paradigm in the research community. [10] utilize the symbolic interface of task planners to construct an abstract state space, facilitating the reuse of hierarchical skills. [25, 57] decompose task specifications into hierarchical neural programs, which feature bottom-level programs as callable subroutines interacting with the environment. [9] chain multiple dexterous policies for achieving long-horizon task goals. [2] adopt a language-based methodology for decomposing action hierarchies. In our work, we introduce an object(affordance)-centric, three-level abstraction framework within OAKINK2 for the decomposition of complex manipulation tasks into *Primitives*.

Motion Synthesis. Motion synthesis involves obtaining credible and realistic human action sequences. There are plenty of works to generate human motions [45, 46, 53], even interactions [15, 33, 34, 51, 52, 56] based on different probabilistic model backbones like cVAE or denoising diffusion. In particular, [33, 34, 52] synthesize human motion based on the object motion, delegating the latter part to preceding models serving as inputs. Inspired by these works, we propose a new task within OAKINK2: Task-aware Motion Fulfillment This task requires the model to synthesize hand motion trajectories based on given textual task descriptions and object motions.

Foundation Models in Manipulation Tasks. Recent days we have seen a significant increase in the application of foundation models in completing manipulation tasks. There are significant efforts for end-to-end foundation models [4, 5, 11] that outputs control signals from visual and textual inputs. Existing works [6, 26, 49] also leverage the in-context learning and zero-shot generalization abilities of Large Language Models (LLMs) for action selection from an array of choices to realize an autoregressive achievement of planning. Demonstration of LLM-based program generation for task completion in [27, 34, 49] inspires us to explore the ability of LLMs to reason code for discerning interdependencies between object affordances in complex tasks, along with the sequence in which they are implemented. Our OAKINK2 introduce the decomposition of *Complex Tasks* into interdependent affordance-based *Primitives*, accompanied by their diverse image-textual descriptions. Based on this, we show an application of OAKINK2 in Complex Task Completion utilizing existing power of foundation models.

3. Construction of OAKINK2

We first introduce how the three-level of abstractions are acquired in Sec. 3.1, then provide the details for data collection and annotation in Sec. 3.2.

3.1. Complex Task Acquisition

Task Initialization. Given a collected repository of objects, we first construct four manipulation scenarios. Each scenario has its unique characteristic and corresponds to a set of complex manipulation tasks. These scenarios are: 1) kitchen table; 2) study room table; 3) demo chem lab; 4) bathroom table. Then, we invite four annotators (👥) to propose *Complex Tasks* in these scenarios and select object cluster that required for these tasks (Fig. 2's 1st column).

3.1.1 Object Affordance Analysis

After the task targets are determined, we proceed to analyze the objects' affordances in given scenarios. The expression of affordance adheres to the definitions in the previous OakInk [60]: each affordance contains a specific object part segmentation (*e.g.* a bottle cap) and a descriptive phrase tuple (*e.g.* `<secure, sth>`), which elucidates the function of that part. We provide examples of these affordances in Fig. 2's 2nd column.

3.1.2 Primitive Task Design

In the second stage, We design *Primitives* as the **minimal** interactions that fulfill those object affordance. Here "minimal" indicates the task are required to fully complete the functionality of a certain affordance without any redundant interaction process. Each *Primitive* contains a starting condition, a terminal condition, and the in-between hand-object interaction process. For example, considering an affordance associated with a knife blade meant to `<cut, sth>`, a corresponding *Primitive*, `cut`, requires the subject to move

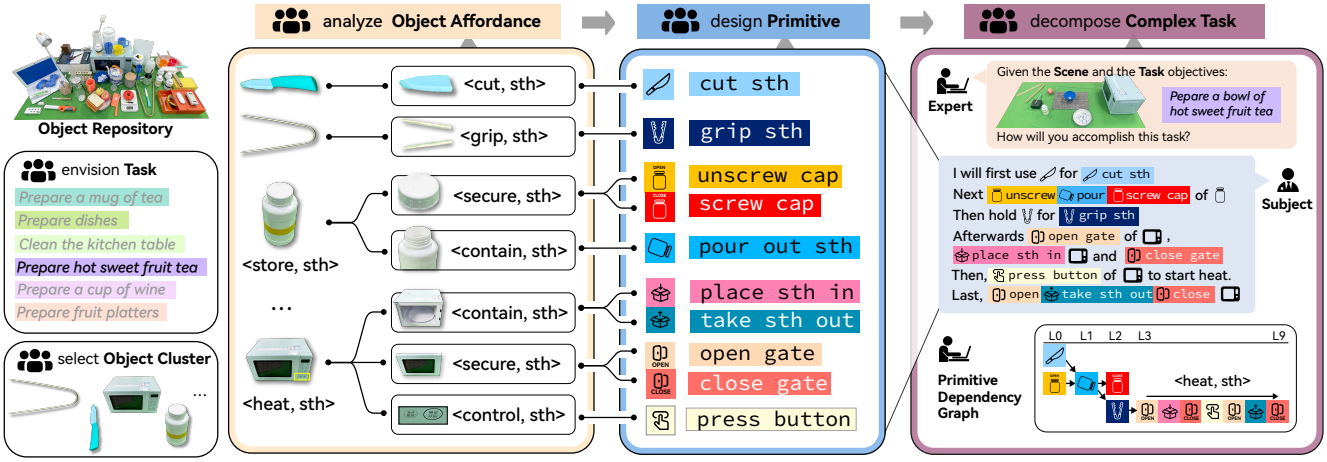

Figure 2. **Illustration of the complex task acquisition process.** This figure use a *Complex Task*: 'Prepare a bowl of hot sweet fruit tea.' to demonstrate the process. Initially, the annotators (👥) analyze the affordances of four essential objects (a gripper, a knife, a tea bottle, and a microwave oven) and design corresponding *Primitive*. For instance, to prepare fruit slices, the *Primitive*: `cut` associated with the knife blade is required. Following this, an expert (🧑‍💻) arranges the scene for the *Complex Task*, and then the subject (🧍), utilizing the designed *Primitive*, plans the execution path of the *Complex Task*. Later, these execution paths are structured into a *Primitive* Dependency Graphs.

the blade to completely pass through the object to be cut so that the separated parts could be detached. In this stage, we collect all available object affordances and their associated *Primitives*, leading to a *Primitive* tasks pool (Fig. 2's 3rd column).

3.1.3 Complex Task Decomposition

In the third stage, we proceed to decompose the previous proposed *Complex Task* – characterized by its long-horizon and multi-goal manipulation targets – into a series of short-term and single-goal *Primitives*. In emphasizing the ordering of *Primitive* completion is important for the *Complex Task* completion, our approach also delineates the **dependencies** between *Primitives*. Therefore, each *Complex Task* contains a series of *Primitives*, along with a *Primitive* Dependency Graph (PDG), which maps out the hierarchical execution order of these *Primitives*. *Primitives* at level 0 (L0) are independent, requiring no prior *Primitives* to be completed, while the final level include those *Primitives* that bring the *Complex Task* to completion.

We deploy a dedicated protocol to acquire the decomposition and dependencies. As shown in Fig. 2's 4th column, initially, an expert (🧑‍💻) instantiates the scene and target with specific description. Subsequently, a subject (🧍) is instructed to describe the order of the completion using the available *Primitive* in the pool. Then, the expert records and organizes this sequence into the PDG, concluding the *Complex Task* acquisition process.

3.2. Data Collection and Annotation

After the acquisition of the three-level of abstractions, the subjects are required to complete the *Primitive* and *Complex Task* respectively in a data capture platform (Fig. 3).

Figure 3. **Capture platform.** 12 MoCap cameras are circled in blue and 4 RGB cameras in red.

3.2.1 Capture Setup

The data capture platform contains two major components: the multi-camera system for recording the manipulation process and the optical MoCap system for pose tracking. The MoCap system uses 12 Optitrack Prime 13W infrared cameras to track the surface markers affixed to the subject's upper body, left and right hand, and interacting objects. The multi-camera system consists of 4 commodity RGB cameras, 3 of which are from allocentric views and 1 is from the egocentric view. We synchronize all sensors at 30 fps and calibrate the transformation between these two systems.

3.2.2 Data Annotation

Object Pose. Poses of rigid bodies are directly solved via the MoCap system. For the poses of articulated bodies, the base parts of articulated bodies are handled similarly to rigid bodies, while the articulated parts are divided into two categories. If the part is large enough to attach enough mark-

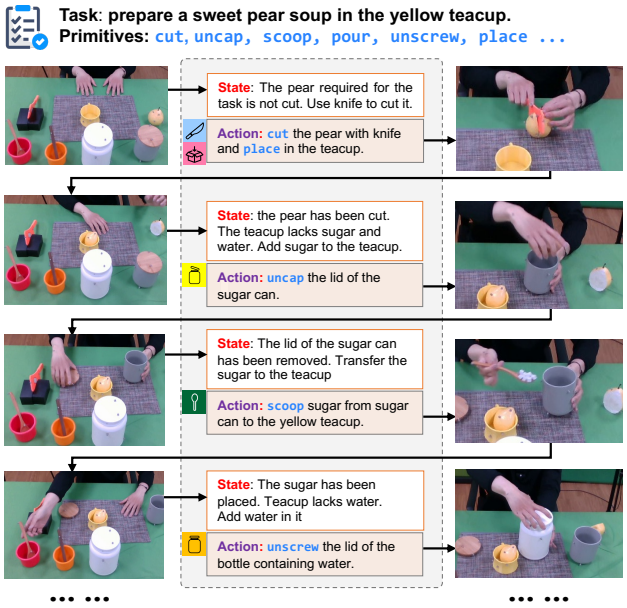

Figure 4. **Commentary of the task execution.** The left column shows the current state of the scene. The center column shows the narrative dialog retrieved from experts. The right column shows the upcoming *Primitive* task to be executed.

ers without blocking the interaction then it will be handled like rigid bodies. Otherwise, only one marker is attached to that part. The marker's position is calibrated in the object's canonical coordinate frame. Later, given the articulation type (*e.g.* revolution or prismatic), the parameter of the articulation joint is determined by minimizing the squared difference between the observed marker position and the recovered marker position in the object's canonical frame.

Human Pose and Surface. The annotation of human pose and surface relies on SMPL-X [44] body mesh. To actually acquire human pose and surface, we employ a two-stage fitting approach in align with the MoSH++ [40]. In the first stage, we use the captured markers when the subject in T-pose to fit the subject's SMPL-X shape parameter $\bar{\beta}$ and each marker's location $P_{\mathcal{M}}^{(c)}$ in SMPL-X canonical space. From stage one's optimization result, we can determine the correspondence $\mathcal{C}(\cdot)$ from the subject's surface markers to the vertices of the SMPL-X model. In the second stage, we fit the per-frame subject poses parameter θ throughout the task completion process. This fitting is grounded in the previously acquired shape $\bar{\beta}$ and marker correspondence $\mathcal{C}(\cdot)$. With pose and shape parameters obtained, the subject's body mesh is reconstructed using the SMPL-X model. Other body representations like MANO are derived from this result. Refer to Sup. Mat for details.

Commentary of Task Execution. After the manipulation process is completed, we send the video recording to experts for analysis, requesting them to furnish detailed commentary on the task execution process. At each *Primitive* step, experts are asked to provide comments on the current task state and the forthcoming action. Specifically, given the execution of the previous *Primitive*, experts are asked to 1) summarize the tasks yet to be completed to achieve the manipulation goals, considering both the current scene and the upcoming *Primitive* slated for execution; and 2) offer descriptions of the next action using the available *Primitives* in the pool. This process is illustrated in Fig. 4. The narrative text provided by experts are subsequently refined using GPT-4 [43] to serve as commentary. OAKINK2 features on these commentaries as they encapsulate the expert's chain-of-thought when observing the manipulation process. These commentaries serve not only to interpret user behaviors but also to inform the generation of user actions.

4. The OAKINK2 Dataset

4.1. Data and Annotation List

OAKINK2 provide RGB videos that record the manipulation processes. These videos are collected from multi-view (1 egocentric and 3 allocentric) setup, synchronized at 30 fps, with resolution 848×480. The annotations contains two parts: **1) 3D motion**, including pose and shape for the human upper-body, hands, and objects (with articulation parameters) during the interaction process; and the **2) task specification**, including object affordances, *Primitives* that correspond to these affordances, *Complex Tasks* with task goals, initial conditions, PDGs, expert commentary, and subject's completion sequence. Evaluations of the 3D annotation qualities are provided in Sup. Mat. Annotation on 3D hand keypoints undergo cross-dataset validation with a reconstruction model, while the 3D poses associated with grasping actions are examined for the physical property integrity.

4.2. Dataset Statistics

OAKINK2 sets up four scenarios of hand-object interaction with a total number of 38 long-horizon complex manipulation goals, which instantiates to 150 *Complex Tasks*. OAKINK2 contains in total 75 objects and 39 affordance. These affordances map to 60 types of *Primitives*. OAKINK2 contains 627 sequences of bimanual dexterous hand-object interaction in total. 363 of these are for *Primitives* and 264 are for *Complex Tasks*. In total, OAKINK2 contains 4.01M image frames. We compare OAKINK2 to multiple existing hand-object interaction datasets in Tab. 1. Here we highlight several notable features of OAKINK2: 1) it provides interaction grounded in object affordance (*vs.* HO3D, DexYCB); 2) it features long-horizon manipulation goals (*vs.* ARCTIC, HOI4D, GRAB); 3) it includes 3D pose and shape annotation for both hands and objects (*vs.* EGO4D, AssemblyHands); and 4) it offers task decomposition using *Primitives*, which is not available in any datasets in Tab. 1.

Dataset	image mod.	resolution	#frame	#views	#subj	#obj	3D gnd.	real / syn.	label method	hand pose	obj pose	afford. inter.	dynamic inter.	long-horizon	task decomp.
EGO4D [17]	✓	∼	∼	1	931	–	✗	–	–	✗	✗	✗	✗	✓	✓
HO3D [19]	✓	640 × 480	78K	1-5	10	10	✓	real	auto	✓	✓	✗	✓	✗	✗
GRAB [50]	✗	–	1.62M	–	10	51	✓	real	mocap	✓	✓	✓	✓	✗	✗
H2O [31]	✓	1280 × 720	571K	5	4	8	✓	real	auto	✓	✓	✓	✓	✗	✗
HOI4D [37]	✓	1280 × 800	3M	1	9	1000	✓	real	crowd	✓	✓	✓	✓	✗	✗
ARCTIC [14]	✓	2800 × 2000	2.1M	9	10	11	✓	real	mocap	✓	✓	✓	✓	✗	✗
AssemblyHands [42]	✓	1920 × 1080	3.03M	12	34	–	✓	real	semi-auto	✓	✗	✓	✓	✓	✓
Ego-Exo4D [18]	✓	∼	∼	5-6	839	–	✓	real	semi-auto	✓	✗	✗	✓	✓	✓
OakInk-Image [60]	✓	848 × 480	230K	4	12	100	✓	real	crowd	✓	✓	✓	✓	✗	✗
OAKINK2	✓	848 × 480	**4.01M**	4	9	75	✓	real	mocap	✓	✓	✓	✓	✓	✓

Table 1. **A cross-comparison among various public datasets.** (Refer to Sup. Mat for the full table.)

5. Selected Applications

5.1. Hand Mesh Reconstruction

The Hand Mesh Reconstruction (HMR) task is to estimate the 3D hand pose during the interaction process from the captured images. We benchmark HMR task under both single-view settings and multi-view settings. In single-view settings, the image input only contains one view, egocentric or allocentric. In multi-view settings, the image input will contain multiple views, together with the camera calibration parameters. For both settings we partition the corresponded task-specified subsets at the sequence level, maintaining the proportion of samples in train/val/test sets at approximately 70%, 5%, and 25%.

We evaluate mean per joint position error (MPJPE), mean per vertex position error (MPVPE) in world space, wrist(root)-relative (RR) systems and systems after Procrustes analysis (PA). We also evaluate area under curve (AUC) of correct keypoints percentage within range $0 - 20\,\mathrm{mm}$ in root-relative systems. We show HMR benchmark results under both settings in Tab. 2.

Setting	Methods	PA-MPJPE	PA-MPVPE	RR-MPJPE (AUC)	RR-MPVPE	MPJPE	MPVPE
Mono	METRO [35]	6.90	6.47	17.56 (0.410)	16.44	–	–
	RLE [32] + HandTailer [39]	5.46	6.86	13.08 (0.441)	14.03	–	–
Multi	KP-based Fit [61]	9.20	8.83	15.63 (0.349)	15.38	19.30	19.11
	POEM [61]	6.18	6.61	12.12 (0.581)	12.15	9.17	9.52

Table 2. Single- and multi-view HMR **evaluation results** in *mm*.

5.2. Task-aware Motion Fulfillment (TaMF)

To achieve task objectives in interaction scenarios, we introduce a novel task: Task-aware Motion Fulfillment (TaMF). It targets at the generation of hand motion sequences that can fulfill given object trajectories conditioned on textual task descriptions.

Task Formulation. Given a textual description of the *Primitive* task: $\mathbf{text}_{\mathrm{PT}}$, we assume the involved objects geometries $\mathcal{V}_o = \{\mathbf{V}_{o,m}\}$ and their motion trajectories $\mathcal{T}_o = \{\mathbf{T}_{o,m}^{(i)}\}$ during the interaction process are known. We use subscript h to represent human hands, o to represent the object, m to index different object instances (and different parts of the same instance) and superscript (i) to index different timestamps. The task is to generate a corresponding hands motion trajectory $\mathcal{P}_h = \{\mathbf{P}_h^{(0:L)}\}$ conditioned on the textual description $\mathbf{text}_{\mathrm{PT}}$, object geometries $\mathcal{V}_o$, and motion trajectories $\mathcal{T}_o$.

Evaluation Metrics. We evaluate contact ratio (CR) and solid intersection volume (SIV) to measure the physical plausibility of the generated motion. On sequence-level, we evaluate motion smoothness with Power Spectrum KL divergence of joints (PSKL-J) as in human motion generation, and evaluate FID to measure distances between the ground-truth motions and the generated motions. We also conduct a perceptual study to evaluate the level of realism for the generated motion. Detailed definition of these metrics can be found in Sup. Mat.

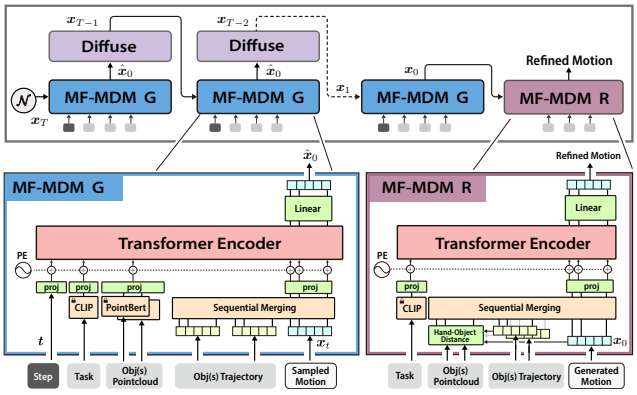

Figure 5. **Architecture of MF-MDM.** First sample random noises $\boldsymbol{x}_T$; then at each step iterating from T to 1, MF-MDM G predicts the cleaned sample $\hat{\boldsymbol{x}}_0$ and then diffuse it back to $\boldsymbol{x}_{t-1}$. After the generated sample $\boldsymbol{x}_0$ is acquired, it is refined by MF-MDM R for better interaction details.

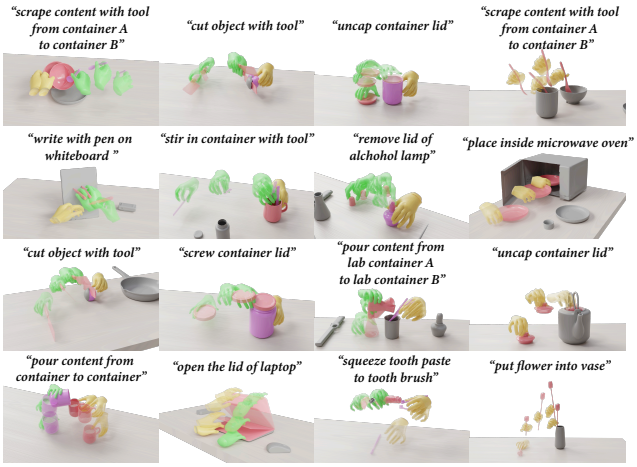

Figure 6. **Qualitative Visualization** of the generated hand motion in TaMF model.

Physical lausibility		Motion Smoothness	
CR↑	SIV (cm^3)↓	PSKL-J (**g.t., p.**) ↓	(**p., g.t.**) ↓
0.90	4.17	0.0446	0.0460

FID	Perceptual Score	
	Dataset	Generated
1.369	4.66 ± 0.48	3.64 ± 0.85

Table 3. **Evaluations** of generated hand motion in TaMF model. PSKL-J is evaluated between the training data (**g.t.**) and the generated hand motion trajectory (**p.**); both directions are included as PSKL-J is an asymmetric metric.

Model and Results. We enhance a diffusion-based motion generation model: MDM [53], tailoring it to the nuanced requirements of task-aware hand motion synthesis. The model architecture is visualized in Fig. 5. Our proposed model, named as MF-MDM, consists of two components: **1) MF-MDM G**, which generates human motion trajectory conditioned on textual descriptions of tasks and object motion trajectories; and **2) MF-MDM R**, which refines generated hand motion based on spatial hand-object relationships. The sampling process is modeled as a reversed diffusion process of gradually cleaning noised samples. The key difference for MF-MDM is to incorporate multi-object related probabilistic conditions into existing transformer encoder. To achieve this, we employ an extra layer, Sequential Merging, to aggregate spatial relationships in the interaction scene at each frame. The object motion trajectories and the previously diffused hand motion trajectory are projected to the same dimension and aggregated. For the refine model MF-MDM R, we append hand-object distances as an extra spatial information for Sequential Merging layer. The aggregated embedding sequence is combined with other tokens before being fed into the main transformer encoder: the noising step token, the text embedding of the task description from the CLIP text encoder, and the aggregated object geometry embeddings from the PointBert encoder. We also provide the quantitative evaluations in Tab. 3 and qualitative visualization in Fig. 6.

5.3. Complex Task Completion (CTC)

OAKINK2 brings in a new application – breaking *Complex Task* goals into paths of *Primitive* motions. The Complex Task Completion (CTC) is to generate hands motion trajectories based on a textual description of the scene and the task objectives. Considering the challenge of direct translation from complex task and scene text to end-to-end motion generation, which involves a transition across multiple modalities, there is currently no adequate framework to address this problem. Therefore, we decompose CTC into three stages, tackling each one sequentially.

The process initially begins with text-based **1) *Primitive* planning**. The recent breakthroughs in foundation models [43, 62], such as Large Language Models (LLMs), allow us to utilize them as the task planner, as these models already have the capability to plan the *Primitive* execution path, while only requiring proper guidance and context. The output of this stage is a task planning script that includes the execution order for each *Primitive*. Subsequently, the problem is reformulated into generating the hand and object motion trajectories for each *Primitive*, based on the target task and scene state, thus modeling $P(\mathcal{P}_h, \mathcal{T}_o | \mathbf{text}_{\mathrm{PT}})$. We again break this down into two subtasks: **2) object trajectory retrieval**, *i.e.* $P(\mathcal{T}_o | \mathbf{text}_{\mathrm{PT}})$ and **3) hand motion generation** *i.e.* $P(\mathcal{P}_h | \mathcal{T}_o, \mathbf{text}_{\mathrm{PT}})$. The former is solved by re-targeting[1] object motion from expert's demonstration to meet the newly generated random scene. The latter is our pre-defined Task-aware Motion Fulfillment model (TaMF, Sec. 5.2).

① Primitive Planning by LLMs. In this stage, we leverage the off-the-shelf GPT-4 [43] to generate program that decompose the *Complex Task* as a sequence of *Primitive*. We first embed the scene description $\mathbf{text}_{\mathrm{scene}}$, the complex task description $\mathbf{text}_{\mathrm{goal}}$ and each object's description $\{\mathbf{text}_{\mathrm{obj}}\}$ into the prompt based on manually designed templates. GPT-4 will respond to the prompt using the **program**. As shown in Fig. 8's code block, this program instantiates the *Primitive* Dependency Graph (PDG) using a sequence of code snippets, where each node of the PDG (*Primitive*) is implemented as a `execute([primitive], ...)` function, and the edge of the PDG is implemented as function's calling order. Then we use a dependency checker built upon the PDG information in OAKINK2 to test whether the generated program completes the *Complex Task* without violation of constraints. If a successful program is obtained, we move to the next stage.

[1] re-target refers to the process of adjusting pre-existing motion trajectories to align with new initial and target poses of objects, ensuring compatibility with the current scene

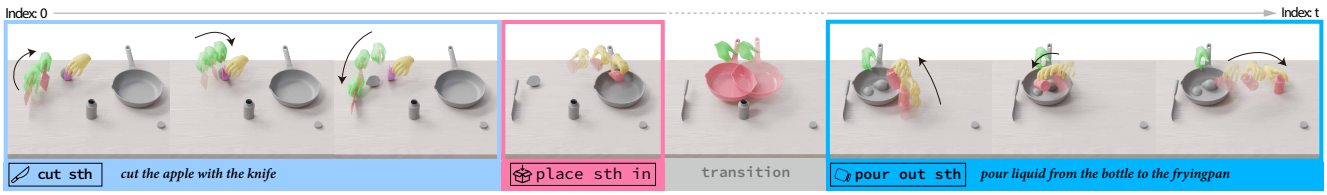

Figure 7. **Visualization of Motion Generation Outcome in Complex Task Completion.**

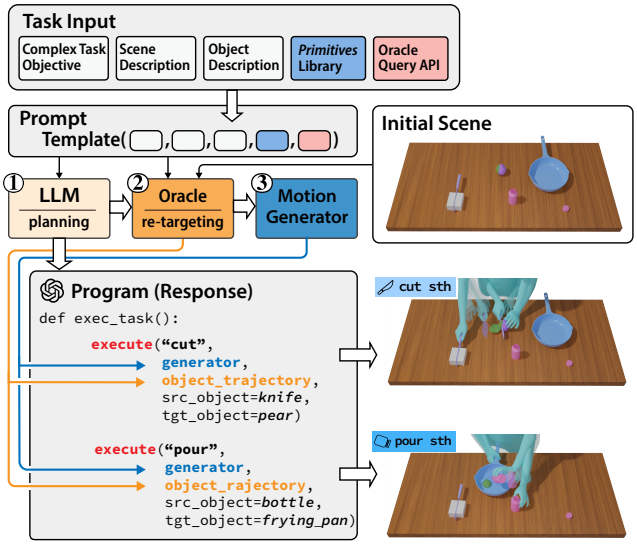

Figure 8. **The diagram of Complex Task Completion.** The task input populates a predefined template to generate the prompt for planning. The ① LLM (GPT-4) responds with code of the program's execution path, delineating the DAG for *Primitive* dependency. Within the code response block, the `orange` snippets marks the ② Oracle to re-target object trajectories; the `blue` snippets indicate ③ motion generators for *Primitives*.

At this moment, the `execute()` function in Fig. 8's code snippets remain incomplete, lacking two pivotal components: the `object_trajectory` and the hand motion `generator`. We will address these components in the following two stages.

② Object Trajectories Retrieval from Oracle. Accomplishing a *Primitive* task necessitates the object's motion trajectories within that context. In this stage, we leverage an Oracle to retrieve object motion trajectories based on a certain scene and *Primitive*. The term "Oracle" denotes a dual-function capability: 1) pursuant to a given *Primitive*, it fetches the object motion trajectories within the OAKINK2 dataset, and 2) it re-targets these expert-derived trajectories based on the initial, functional and post poses of the objects, thereby conforming to new scene requirements and generating the desired `object_trajectory`.

③ Hand Motion Generation with TaMF. Once the object trajectories are obtained, the final stage is to generate hand motion trajectories for each *Primitive*. To this end, we utilize our previously designed Task-aware Motion Fulfillment model (TaMF, Sec. 5.2) as a generalist `generator` (indicating that a singular TaMF model accommodates all *Primitives*). After populating all `execute()` functions with the determined object trajectories and generator, the program is executed in sequel and all the *Primitive* trajectories are connected by interpolation. This interpolation ensures smooth transitions by linking the final state of a preceding trajectory with the initial state of the subsequent one.

We show an example of the generated motions for *Complex Task* in Fig. 7. Details of test scene generation, prompts and templates, evaluations of primitive planning, success/-failure cases are referred to Sup. Mat.

6. Future Works

OAKINK2 is a dataset packing a variety of hand-object interactions for human completion of long-horizon and multi-goal complex manipulation tasks. OAKINK2 incorporates *Primitive* demonstrations, characterized as minimal interactions that fulfill object affordance, and *Complex Tasks* demonstrations, which also include their decomposition into interdependent *Primitives*.

First, we expect OAKINK2 to support large-scale language-manipulation pre-training, improving the performance of multi-modal (*e.g.* vision-language-action [62]) models for Complex Task Completion. In the longer term, we expect OAKINK2 can potentially support learning frameworks capable of end-to-end text-to-manipulation generation.

Second, OAKINK2 can empower various embodied manipulation tasks by re-targeting the collected demonstrations of *Primitives* to different embodiments, such as heterogeneous hands and platforms as [21, 47, 48, 55, 58] implied. The interaction scenarios constructed in OAKINK2 can also be transferred and integrated into existing simulation environments [41, 54] to support embodied learning on object manipulation.

Acknowledgments. This work was supported by the National Key Research and Development Project of China (No. 2022ZD0160102), National Key Research and Development Project of China (No. 2021ZD0110704), Shanghai Artificial Intelligence Laboratory, XPLORER PRIZE grants, and 2023 Shanghai Pujiang X Program Project (No. 23511103104).

References

[1] Ahmed Tawfik Aboukhadra, Jameel Malik, Ahmed Elhayek, Nadia Robertini, and Didier Stricker. THOR-Net: End-to-end graformer-based realistic two hands and object reconstruction with self-supervision. In *Winter Conference on Applications of Computer Vision (WACV)*, 2023. 2

[2] Suneel Belkhale, Tianli Ding, Ted Xiao, Pierre Sermanet, Quon Vuong, Jonathan Tompson, Yevgen Chebotar, Debidatta Dwibedi, and Dorsa Sadigh. RT-H: Action hierarchies using language. *arXiv preprint arXiv:2403.01823*, 2024. 3

[3] 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 *European Conference on Computer Vision (ECCV)*, 2020. 2, 3

[4] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alex Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. 3

[5] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-1: Robotics transformer for real-world control at scale. *Robotics: Science and Systems (RSS)*, 2023. 3

[6] Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on Robot Learning (CoRL)*, 2023. 3

[7] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[8] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3

[9] Yuanpei Chen, Chen Wang, Li Fei-Fei, and C Karen Liu. Sequential dexterity: Chaining dexterous policies for long-horizon manipulation. *arXiv preprint arXiv:2309.00987*, 2023. 3

[10] Shuo Cheng and Danfei Xu. LEAGUE: Guided skill learning and abstraction for long-horizon manipulation. *IEEE Robotics and Automation Letters*, 2023. 3

[11] Open X-Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, Andrew Wang, Anikait Singh, Animesh Garg, Aniruddha Kembhavi, Annie Xie, Anthony Brohan, Antonin Raffin, Archit Sharma, Arefeh Yavary, Arhan Jain, Ashwin Balakrishna, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Blake Wulfe, Brian Ichter, Cewu Lu, Charles Xu, Charlotte Le, Chelsea Finn, Chen Wang, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Christopher Agia, Chuer Pan, Chuyuan Fu, Coline Devin, Danfei Xu, Daniel Morton, Danny Driess, Daphne Chen, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dinesh Jayaraman, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Ethan Foster, Fangchen Liu, Federico Ceola, Fei Xia, Feiyu Zhao, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Gilbert Feng, Giulio Schiavi, Glen Berseth, Gregory Kahn, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homer Walke, Hongjie Fang, Huy Ha, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jaimyn Drake, Jan Peters, Jan Schneider, Jasmine Hsu, Jeannette Bohg, Jeffrey Bingham, Jeffrey Wu, Jensen Gao, Jiaheng Hu, Jiajun Wu, Jialin Wu, Jiankai Sun, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jimmy Wu, Jingpei Lu, Jingyun Yang, Jitendra Malik, João Silvério, Joey Hejna, Jonathan Booher, Jonathan Tompson, Jonathan Yang, Jordi Salvador, Joseph J. Lim, Junhyek Han, Kaiyuan Wang, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Black, Kevin Lin, Kevin Zhang, Kiana Ehsani, Kiran Lekkala, Kirsty Ellis, Krishan Rana, Krishnan Srinivasan, Kuan Fang, Kunal Pratap Singh, Kuo-Hao Zeng, Kyle Hatch, Kyle Hsu, Laurent Itti, Lawrence Yunliang Chen, Lerrel Pinto, Li Fei-Fei, Liam Tan, Linxi "Jim" Fan, Lionel Ott, Lisa Lee, Luca Weihs, Magnum Chen, Marion Lepert, Marius Memmel, Masayoshi Tomizuka, Masha Itkina, Mateo Guaman Castro, Max Spero, Maximilian Du, Michael Ahn, Michael C. Yip, Mingtong Zhang, Mingyu Ding, Minho Heo, Mo-

han Kumar Srirama, Mohit Sharma, Moo Jin Kim, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Ning Liu, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Osbert Bastani, Pannag R Sanketi, Patrick "Tree" Miller, Patrick Yin, Paul Wohlhart, Peng Xu, Peter David Fagan, Peter Mitrano, Pierre Sermanet, Pieter Abbeel, Priya Sundaresan, Qiuyu Chen, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Mart'in-Mart'in, Rohan Baijal, Rosario Scalise, Rose Hendrix, Roy Lin, Runjia Qian, Ruohan Zhang, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Shan Lin, Sherry Moore, Shikhar Bahl, Shivin Dass, Shubham Sonawani, Shuran Song, Sichun Xu, Siddhant Haldar, Siddharth Karamcheti, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Subramanian Ramamoorthy, Sudeep Dasari, Suneel Belkhale, Sungjae Park, Suraj Nair, Suvir Mirchandani, Takayuki Osa, Tanmay Gupta, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Thomas Kollar, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Trinity Chung, Vidhi Jain, Vincent Vanhoucke, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiaolong Wang, Xinghao Zhu, Xinyang Geng, Xiyuan Liu, Xu Liangwei, Xuanlin Li, Yao Lu, Yecheng Jason Ma, Yejin Kim, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Yilin Wu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yue Cao, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunchu Zhang, Yunfan Jiang, Yunshuang Li, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zehan Ma, Zhuo Xu, Zichen Jeff Cui, Zichen Zhang, and Zipeng Lin. Open X-Embodiment: Robotic learning datasets and RT-X models. https://arxiv.org/abs/2310.08864, 2023. 3

[12] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. GanHand: Predicting human grasp affordances in multi-object scenes. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3

[13] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David Crandall. HOPE-Net: A graph-based model for hand-object pose estimation. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[14] 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 *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 6

[15] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. IMoS: Intent-driven full-body motion synthesis for human-object interactions. In *Computer Graphics Forum*, 2023. 2, 3

[16] James J Gibson. *The ecological approach to visual perception: classic edition*. Psychology Press, 2014. 2

[17] 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 *Computer Vision and Pattern Recognition (CVPR)*, 2022. 6

[18] 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. *arXiv preprint arXiv:2311.18259*, 2023. 6

[19] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3D annotation of hand and object poses. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3, 6

[20] 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 *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[21] Ankur Handa, Karl Van Wyk, Wei Yang, Jacky Liang, Yu-Wei Chao, Qian Wan, Stan Birchfield, Nathan Ratliff, and Dieter Fox. DexPilot: Vision-based teleoperation of dexterous robotic hand-arm system. In *International Conference on Robotics and Automation (ICRA)*, pages 9164–9170. IEEE, 2020. 8

[22] 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 *Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3

[23] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, 2020.

[24] Yana Hasson, Gül Varol, Ivan Laptev, and Cordelia Schmid. Towards unconstrained joint hand-object reconstruction from rgb videos. In *International Conference on 3D Vision (3DV)*, 2021. 2

[25] De-An Huang, Suraj Nair, Danfei Xu, Yuke Zhu, Animesh Garg, Li Fei-Fei, Silvio Savarese, and Juan Carlos Niebles. Neural task graphs: Generalizing to unseen tasks from a single video demonstration. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[26] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning (ICML)*. PMLR, 2022. 3

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

[28] 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. *arXiv preprint arXiv:2309.08942*, 2023. 2, 3

[29] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[30] Christopher A Kurby and Jeffrey M Zacks. Segmentation in the perception and memory of events. *Trends in cognitive sciences*, 2008. 2

[31] Taein Kwon, Bugra Tekin, Jan Stuhmer, Federica Bogo, and Marc Pollefeys. H2O: Two hands manipulating objects for first person interaction recognition. In *International Conference on Computer Vision (ICCV)*, 2021. 2, 3, 6

[32] Jiefeng Li, Siyuan Bian, Ailing Zeng, Can Wang, Bo Pang, Wentao Liu, and Cewu Lu. Human pose regression with residual log-likelihood estimation. In *International Conference on Computer Vision (ICCV)*, 2021. 6

[33] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 2023. 3

[34] Kailin Li, Lixin Yang, Zenan Lin, Jian Xu, Xinyu Zhan, Yifei Zhao, Pengxiang Zhu, Wenxiong Kang, Kejian Wu, and Cewu Lu. FAVOR: Full-body ar-driven virtual object rearrangement guided by instruction text. In *AAAI Conference on Artificial Intelligence*, 2024. 3

[35] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 6

[36] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3D hand-object poses estimation with interactions in time. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[37] 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 *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 6

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

[39] Jun Lv, Wenqiang Xu, Lixin Yang, Sucheng Qian, Chongzhao Mao, and Cewu Lu. HandTailor: Towards high-precision monocular 3D hand recovery. In *British Machine Vision Conference (BMVC)*, 2021. 6

[40] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. AMASS: Archive of motion capture as surface shapes. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[41] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021. 8

[42] Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. AssemblyHands: Towards egocentric activity understanding via 3D hand pose estimation. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 6

[43] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 5, 7

[44] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive Body Capture: 3D hands, face, and body from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[45] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[46] Mathis Petrovich, Michael J Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[47] Yuzhe Qin, Hao Su, and Xiaolong Wang. From one hand to multiple hands: Imitation learning for dexterous manipulation from single-camera teleoperation. *IEEE Robotics and Automation Letters*, 2022. 2, 8

[48] Yuzhe Qin, Yueh-Hua Wu, Shaowei Liu, Hanwen Jiang, Ruihan Yang, Yang Fu, and Xiaolong Wang. DexMV: Imitation learning for dexterous manipulation from human videos. In *European Conference on Computer Vision (ECCV)*, 2022. 2, 8

[49] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. ProgPrompt: Generating situated robot task plans using large language models. In *International Conference on Robotics and Automation (ICRA)*, 2023. 3

[50] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 3, 6

[51] Omid Taheri, Vasileios Choutas, Michael J Black, and Dimitrios Tzionas. GOAL: Generating 4d whole-body motion for hand-object grasping. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3

[52] Omid Taheri, Yi Zhou, Dimitrios Tzionas, Yang Zhou, Duygu Ceylan, Soren Pirk, and Michael J Black. Grip: Generating interaction poses using spatial cues and latent consistency. In *International Conference on 3D Vision (3DV)*, 2024. 3

[53] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *International Conference on Learning Representations (ICLR)*, 2023. 3, 7

[54] Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012. 8

[55] Weikang Wan, Haoran Geng, Yun Liu, Zikang Shan, Yaodong Yang, Li Yi, and He Wang. UniDexGrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning. *arXiv preprint arXiv:2304.00464*, 2023. 8

[56] Yan Wu, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang. SAGA: Stochastic whole-body grasping with contact. In *European Conference on Computer Vision (ECCV)*, 2022. 2, 3

[57] Danfei Xu, Suraj Nair, Yuke Zhu, Julian Gao, Animesh Garg, Li Fei-Fei, and Silvio Savarese. Neural task programming:

Learning to generalize across hierarchical tasks. In *International Conference on Robotics and Automation (ICRA)*, 2018. 3

[58] Yinzhen Xu, Weikang Wan, Jialiang Zhang, Haoran Liu, Zikang Shan, Hao Shen, Ruicheng Wang, Haoran Geng, Yijia Weng, Jiayi Chen, et al. Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 8

[59] Lixin Yang, Kailin Li, Xinyu Zhan, Jun Lv, Wenqiang Xu, Jiefeng Li, and Cewu Lu. ArtiBoost: Boosting articulated 3D hand-object pose estimation via online exploration and synthesis. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[60] 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 *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 6

[61] Lixin Yang, Jian Xu, Licheng Zhong, Xinyu Zhan, Zhicheng Wang, Kejian Wu, and Cewu Lu. POEM: Reconstructing hand in a point embedded multi-view stereo. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 6

[62] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3D-VLA: A 3D vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*, 2024. 7, 8

[63] Juntian Zheng, Qingyuan Zheng, Lixing Fang, Yun Liu, and Li Yi. CAMS: Canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[64] 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. *arXiv preprint arXiv:2305.01618*, 2023. 2, 3