TASTE-Rob: Advancing Video Generation of Task-Oriented Hand-Object Interaction for Generalizable Robotic Manipulation

Hongxiang Zhao[1*] Xingchen Liu[1*†] Mutian Xu[1] Yiming Hao[1] Weikai Chen[‡] Xiaoguang Han[1,2§]

[1]SSE, CUHKSZ [2]FNii, CUHKSZ
https://taste-rob.github.io

Abstract

We address key limitations in existing datasets and models for task-oriented hand-object interaction video generation, a critical approach of generating video demonstrations for robotic imitation learning. Current datasets, such as Ego4D [16], often suffer from inconsistent view perspectives and misaligned interactions, leading to reduced video quality and limiting their applicability for precise imitation learning tasks. Towards this end, we introduce TASTE-Rob — a pioneering large-scale dataset of 100,856 ego-centric hand-object interaction videos. Each video is meticulously aligned with language instructions and recorded from a consistent camera viewpoint to ensure interaction clarity. By fine-tuning a Video Diffusion Model (VDM) on TASTE-Rob, we achieve realistic object interactions, though we observed occasional inconsistencies in hand grasping postures. To enhance realism, we introduce a three-stage pose-refinement pipeline that improves hand posture accuracy in generated videos. Our curated dataset, coupled with the specialized pose-refinement framework, provides notable performance gains in generating high-quality, task-oriented hand-object interaction videos, resulting in achieving superior generalizable robotic manipulation. To foster further advancements in the field, TASTE-Rob dataset and source code will be made publicly available on our website https://taste-rob.github.io.

1. Introduction

Robotic manipulation plays an essential role in daily life, assisting with tasks like object handling, liquid pouring, surface cleaning, and drawer operation. Powered by the success of Imitation Learning (IL) [8–10, 15, 20, 22, 42, 48, 56, 57, 64], current robots can imitate actions from video demonstrations but require a nearly identical environment in the recorded video, limiting their generalization to novel scenes where factors like the position or movement direction of the manipulated object differ.

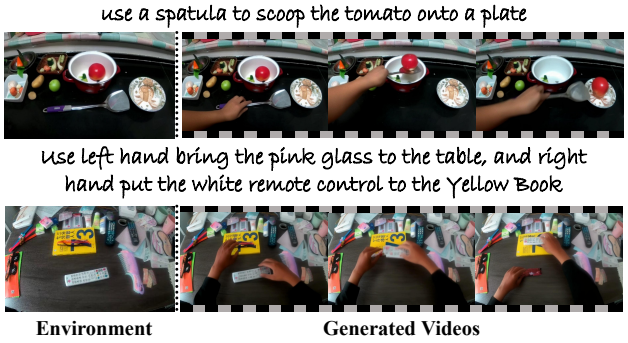

Figure 1. **TASTE-Rob** generates high-quality task-oriented hand-object interaction videos, enabling robust generalization in robotic manipulation.

To solve the issue of limited generalization, recent research combines IL with generative models [2, 3, 5–7, 12, 25, 47, 58], like video diffusion models (VDMs), to create adaptable demonstrations and enable more generalizable robotic manipulation [12, 25, 47]. Specifically, these works develop VDMs to generate robot-object interaction videos as demonstrations for IL. While robot-object interaction VDMs enhance generalization performance, they are constrained by limited robot datasets [4, 8, 21, 24, 51]. A scalable alternative is generating hand-object interaction (HOI) videos, which leverage extensive human manipulation video datasets. Although hand-object interaction differs from robotic motion, advanced policy models [1, 5, 9, 22, 48, 57, 65] can bridge this gap effectively, achieving remarkable performance through IL. However, even if existing powerful VDMs are able to generate remarkable human videos, as shown in Figure 7, they struggle with generating high-quality, task-oriented hand-object interaction videos with accurate task understanding.

*Equal contribution.
†Work done as a visiting researcher at CUHKSZ.
‡This paper solely reflects the author's personal research and is not associated with the author's affiliated institution.
§Corresponding author: hanxiaoguang@cuhk.edu.cn.

To enhance video generation quality, Ego4D [16], a large-scale data containing 83,647 ego-centric hand-object interaction videos, has proven useful in robotic manipulation tasks [23, 32, 53]. However, Ego4D presents two notable limitations for our purpose, as shown in Appendix 8. Firstly, the camera perspective in each clip is inconsistent, whereas video demonstrations for IL are typically recorded from a fixed viewpoint. Secondly, since these clips are extracted from longer videos that encompass multiple tasks, some clips contain partial interactions that overlap with others, resulting in misalignment between video clips and language instructions. Both limitations negatively impact the quality of video generation.

To address these above limitations, we introduce TASTE-Rob, a pioneering large-scale dataset designed to enhance the generation quality of task-oriented hand-object interaction videos. TASTE-Rob contains 100,856 ego-centric videos with well-aligned language instructions. Unlike Ego4D, TASTE-Rob is crafted to ensure clarity and consistency – each video captures a single, complete interaction corresponding to its instruction and is recorded independently with a fixed camera viewpoint, mirroring the conditions needed for reliable imitation learning demonstrations. By resolving these key issues, TASTE-Rob provides a robust foundation for advancing research in video generation, imitation learning, and related fields.

After fine-tuning a VDM on our curated dataset, we achieve high-quality video generation of task-oriented hand-object interactions. While the model demonstrates proficiency in identifying target objects and determining appropriate coarse interactions, it exhibits inconsistencies in hand poses, with the grip changing unnaturally during manipulation. This limitation presents a critical challenge for robotic imitation learning where precise and stable hand-object interactions are essential. To address this issue, we further propose a dedicated three-stage pipeline for pose refinement. First, we generate a coarse manipulation video conditioned solely on language instructions and environment images. Second, we refine the hand pose sequences using a Motion Diffusion Model (MDM) [45], ensuring realistic grip poses. Finally, we regenerate the videos by incorporating the revised pose sequences as an additional conditioning factor. This pipeline enhances grasping realism in generated videos, enabling robots to execute novel tasks with greater accuracy and adaptability in unseen scenarios. Extensive experiments demonstrate that our proposed TASTE-Rob dataset, combined with the pose-refinement pipeline, achieves significant performance gains, surpassing state-of-the-art methods in the quality of generated hand-object interaction videos and achieving accurate robotic manipulation.

Our contributions can be summarized as follows:

- We introduce TASTE-Rob, a large-scale dataset containing 100,856 ego-centric hand-object interaction videos, each specially tailored for training high-quality video generation model for robot imitation learning.
- We develop a three-stage pose-refinement pipeline that significantly enhances the fidelity of the hand posture in the generated videos.
- We set a new state of the art in generating task-oriented hand-object interaction videos, which further boosts the performance of robot manipulation.

2. Related Works

Imitation Learning from video demonstrations. IL enables robots to learn from expert [8, 10, 15, 42, 64] or human [9, 20, 22, 48, 56, 57] video demonstrations by training policies to directly replicate expert actions through supervised learning. While these methods require video demonstrations from environments identical to the robot execution setup, they exhibit limited generalization to novel scenes and tasks. Alternatively, existing works address this challenge by generating demonstrations, such as videos [5, 23, 32, 53], optical flows [52, 58], and images [7], successfully achieving better generalization. Among existing approaches, video demonstration generation represents the most straightforward solution. Yet, robot manipulation video generation [23, 32, 53] is constrained by limited robot datasets, while human hand video generation [5] is hindered by the absence of effective HOI video generative models. In this work, we explore enhancing HOI video generation quality to facilitate robot manipulation.

Ego-Centric HOI Video datasets. To enhance video quality, we propose a specialized HOI video generation model, which requires a large-scale, high-quality ego-centric dataset. Existing ego-centric HOI datasets have various limitations: insufficient resolution [27, 28, 30, 43], limited scale [13, 27, 28, 30, 35, 43], and restricted scene diversity [11, 13, 28]. While Ego4D [16] provides diverse, large-scale HOI videos, its varied camera perspectives and imprecise action-language alignment pose challenges for IL video demonstrations. Therefore, we introduce TASTE-Rob, containing 100,856 videos specifically designed for our task.

Diffusion Models for Video Generation. Based on the success of VDMs [18], researchers have achieved high-quality video generation with various conditioning inputs: text [33, 41, 61, 62] and images [26, 38, 54, 55, 60]. Image-to-video generation aligns well with our task. While these models excel at producing realistic colors and temporal coherence, they struggle with significant motion variations. Motion diffusion models [19, 36, 45, 50, 59] address this limitation by leveraging high-quality human body models [39] to generate motion control parameters. In this work,

Orientation(°)	0-90	90-180	180-270	270-360
Proportion(%)	33.82	47.7	9.59	8.89

Table 1. **Palm orientation distribution of grasping poses.** We analyze the palm orientation distribution of grasping poses across TASTE-Rob. Specifically, the degree of orientation denotes the angle between the palm normal direction and the positive x-axis of the frame plane.

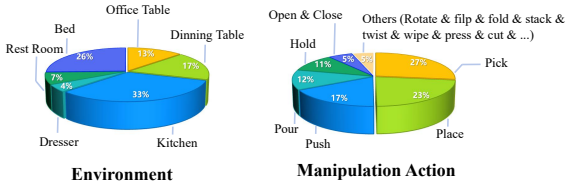

Figure 2. **Distribution of environment and HOI task.** In TASTE-Rob, we collect ego-centric task-oriented HOI videos with diverse tasks under diverse environments.

we combine the advantages of VDM and MDM through our pose-refinement pipeline to generate HOI videos with realistic grasping poses.

3. Building TASTE-Rob Dataset

We collect a massive and diverse ego-centric task-oriented HOI video dataset **TASTE-Rob**, which includes 100,856 pairs of video and their corresponding language task instruction. To serve for HOI video generation, TASTE-Rob is required to achieve following goals: 1) Each video is recorded with a static camera perspective and contains a single action that closely aligns with the task instruction. 2) It encompasses diversified environments and tasks. 3) It features a variety of hand poses across different HOI scenarios.

3.1. Collection Strategy and Camera Setting

To achieve the first goal, we utilize multiple cameras equipped with a wide-angle lens capable of capturing 1080p ego-centric videos. We make the following improvements during each recording process.

Firstly, since our data collection aims to generate task-oriented HOI videos for demonstrations in IL, where demonstrations are typically recorded from fixed camera viewpoints for effective robot imitation learning, we ensure that no camera perspective variation occurs during recording. In addition, as shown in Figure 12, we align the camera perspective specifically to match the head-mounted camera setup of Ego4D [16], ensuring consistency with the ego-centric perspective.

Our second objective is to ensure precise alignment between language task instructions and video actions in TASTE-Rob, a crucial aspect for maintaining the action integrity in generated HOI videos. Unlike Ego4D [16], which captures extended recordings of daily activities via head-mounted cameras and later segments them into shorter clips,

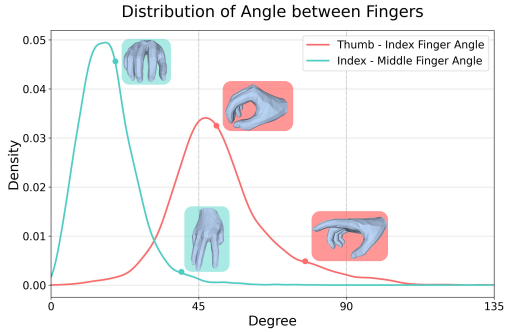

Figure 3. **Inter-finger angle distribution of human hands.** We analyze the distribution of inter-finger angles, specifically between the thumb-index and index-middle finger pairs, across TASTE-Rob. Specifically, higher degrees mean higher inter-finger angles.

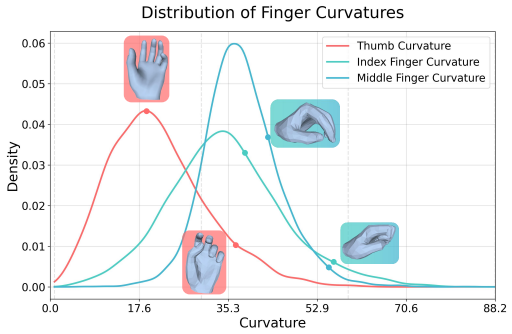

Figure 4. **Finger curvature distribution of human hands.** We analyze the distribution of finger curvatures, specifically that of thumb, index finger and middle finger, across TASTE-Rob. Specifically, higher degrees mean higher curvature.

we employ a more controlled collection protocol: 1) Each video is strictly limited to under eight seconds in duration and captures a single action. 2) Collectors follow a structured recording process: they press the "start recording" button, execute the specified HOI task according to the provided instruction, and stop recording upon task completion. This approach ensures a precise correspondence between actions and task instructions.

3.2. Data Diversity

Distribution of environment and task. To promote broad generalization, videos in TASTE-Rob are recorded across diverse environments and cover a wide range of HOI tasks. As shown in Figure 2, environments include locations such as kitchens, bedrooms, dinning tables, office tables, etc. Collectors are required to interact with various commonly used objects, and perform tasks including picking, placing, pushing, pouring, etc. To further ensure task diversity, we consider different patterns of hand usage. Specifically, TASTE-Rob comprises 75,389 single-hand task videos and 25,467 double-hand task videos.

Dataset	Language Instructions		Environment		Data Statistics		
	Perfect Action Alignment	Determiner	Camera	Setting	Frames	Clips	Resolution
Disney [13]	✗	✗	Dynamic	Garden	11M	8800	1080p
ADL [35]	✗	✗	Dynamic	Diverse	1M	436	960p
Charades-Ego [43]	✗	✗	Dynamic	Diverse	5.8M	7860	720p
HOI4D [29]	✓	✗	Static	Diverse	4M	2466	1080p
UT Ego [27, 30]	✗	✗	Dynamic	Diverse	0.9M	80	320p
EGTEA Gaze+ [28]	✓	✗	Dynamic	Kitchen	2.5M	10,000	480p
EPIC-KITCHENS [11]	✓	✗	Dynamic	Kitchen	6.8M	35,000	1080p
Ego4D [16]	✗	✗	Dynamic	Diverse	19.2M	83,647	1080p
Ours	✓	✓	Static	Diverse	9M	100,856	1080p

Table 2. **Dataset comparison with existing ego-centric HOI video datasets.** TASTE-Rob, specifically designed for generating task-oriented HOI videos, provides a diverse collection of ego-centric HOI videos with several distinctive features: static camera perspectives, precise language-action alignment, and rich object descriptors (Please see Appendix 8 for details). In the above table, we compare these key characteristics: Perfect Action Alignment (ensuring all actions in the video strictly correspond to language task descriptions), Descriptive Determiners (whether object descriptions exist), Camera Setting (fixed or dynamic camera perspective), and Environment Setting (recording locations). By the way, Even if Ego4D [16] offers an impressive 3000-hours video data, it is designed for various general tasks, and the portion applicable to HOI video generation is significantly smaller, as presented in the above table, which is only 83,647 clips. Cells highlighted in denotes the dataset with best feature in each column.

Distribution of Grasping Hands. To ensure a wide variety of hand poses, we consider two primary factors: diverse palm orientations (global pose) and varied grasping poses (detailed pose). To demonstrate the diversity of hand poses, we utilize HaMeR [34] to extract hand pose parameters and analyze the distribution based on these parameters (see further details in Appendix 9).

As shown in Table 1, we analyze the distribution of palm orientations during HOI interaction across TASTE-Rob. The analysis reveals the following insights: 1) Hand poses with palms facing down (0°-180°) are the most common, as this orientation is practical for grasping objects. 2) Hand poses with palms facing left (90°-270°) occur slightly more frequently than those facing right, likely because all collectors are right-handed and naturally prefer using their right hand for object manipulation.

We also provide the analysis of hand grasping pose distribution in Figures 3 and 4. Given the dominant roles of the thumb, index finger, and middle finger in HOI, we focus on examining both the inter-finger angles between these fingers and their individual curvature distributions. The analysis in Figure 3 shows a broad distribution of inter-finger angles, indicating diverse hand orientations. Figure 4 reveals two key findings: 1) The curvature distributions of the index and middle fingers exhibit similar patterns, reflecting their synchronized bending during HOI actions. 2) Our dataset captures a broad spectrum of grasping poses, driven by the variety of objects being manipulated.

3.3. Comparisons with Other Datasets

As shown in Table 2, we provide comparison between TASTE-Rob and existing ego-centric HOI video datasets. TASTE-Rob is the first video dataset specifically designed for task-oriented HOI video generation, and it can also serve as a valuable resource for demonstrations in IL. Given that IL video demonstrations are recorded from a fixed camera perspective and include only a single action that aligns with the task instruction, we collect HOI videos under the same settings, distinguishing TASTE-Rob from other datasets. In addition, to improve comprehension of target objects, we incorporate diverse determiners of objects in language task instructions (see Appendix 8 for further details). With TASTE-Rob, we are able to generate high-quality HOI video demonstrations to achieve IL.

4. Method

Given an environment image and a task description, the generated task-oriented HOI videos are required to satisfy: 1) **Accurate Task Understanding**: correctly identifying which object to manipulate and how to manipulate it. 2) **Feasible HOI**: maintaining consistent hand grasping poses throughout the manipulation.

As illustrated in **Stage I** area of Figure 5, while videos generated by a single VDM ($\hat{v}_c$) demonstrate accurate task understanding, exhibit limited fidelity in maintaining consistent grasping poses. To fulfill both requirements, we propose a three-stage pose-refinement pipeline, outlined in Figure 5: **Stage I**: employing a learnable Image-to-Video (I2V) diffusion model to generate a coarse HOI video that meets "Accurate Task Understanding". **Stage II**: extracting a hand pose sequence from this coarse video and refining it using a learnable Motion Diffusion Model (MDM) [45]. **Stage III**: using the refined hand pose sequence to generate a high-

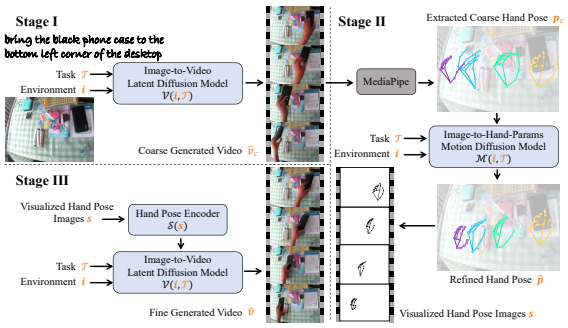

Figure 5. **Overview of our proposed three-stage pose-refinement pipeline. Stage I**: our learnable I2V latent diffusion model $\mathcal{V}$ generates a coarse HOI video $\hat{\boldsymbol{v}}_c$. **Stage II**: extracting hand pose sequence $\boldsymbol{p}_c$ from $\hat{\boldsymbol{v}}_c$, and our learnable MDM $\mathcal{M}$ refines it into $\hat{\boldsymbol{p}}$. Besides, we also obtain the corresponding visualized hand pose images $\boldsymbol{s}$ of $\hat{\boldsymbol{p}}$. **Stage III**: incorporating this additional pose condition into $\mathcal{V}$ to generate a fine HOI video $\hat{\boldsymbol{v}}$.

fidelity HOI video that satisfies both requirements.

4.1. Preliminary: Diffusion Models

Diffusion models [17, 44] work by learning to reverse a noising process. They firstly transform data samples $\boldsymbol{x}_0 \sim p_{data}(\boldsymbol{x})$ into Gaussian noise $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, then learn to reconstruct the original data by denoising $\boldsymbol{x}_T$ with multiple steps. With the following objective:

$$\min_{\theta} \mathbb{E}_{t,\boldsymbol{x},\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \|\epsilon - \epsilon_\theta(\boldsymbol{x}_t, t)\|_2^2, \tag{1}$$

the model $\epsilon_\theta(\boldsymbol{x}_t, t)$ learns to predict single-step noise at each timestep t, where ϵ is the sampled ground truth noise and θ indicates the learnable network parameters. After training, the model generates $\hat{\boldsymbol{x}}_0$ through a T-step denoising process from a random Gaussian noise $\boldsymbol{x}_T$.

Video Generation. In this work, we explore HOI video generation based on DynamiCrafter [54], a powerful I2V latent diffusion model. Suppose $\mathcal{T}$, $\boldsymbol{i} \in \mathbb{R}^{3\times H\times W}$, and $\boldsymbol{v} \in \mathbb{R}^{L\times 3\times H\times W}$ denote task language descriptions, environment images, and ground truth video frames, respectively. DynamiCrafter learns the denoising process in a compact latent space: $\boldsymbol{v}$ is encoded into a compact latent space through an encoder $\mathcal{E}$, yielding latent representations $\boldsymbol{z} = \mathcal{E}(\boldsymbol{v}) \in \mathbb{R}^{L\times C\times h\times w}$, and decoded via the decoder $\mathcal{D}$. Within this latent space, the model performs both forward diffusion and guided denoising processes conditioned on $\mathcal{T}$ and $\boldsymbol{i}$. Specifically, $\boldsymbol{z}_0 \sim p_{data}(\boldsymbol{z})$ and $\boldsymbol{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The optimization objective 1 is updated as:

$$\min_{\theta_V} \mathbb{E}_{t,\boldsymbol{z},\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \|\epsilon - \epsilon_\theta(\boldsymbol{z}_t, \mathcal{T}, \boldsymbol{i}, t, fps)\|_2^2, \tag{2}$$

where fps is the FPS control introduced in [54] and θ_V denotes the learnable parameters of DynamiCrafter.

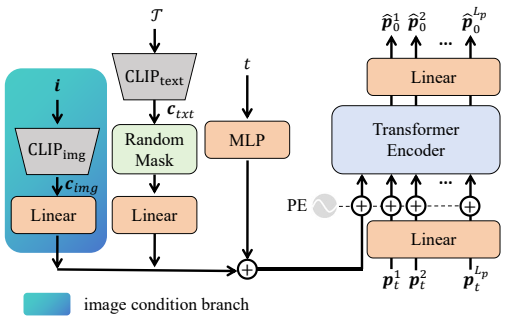

Figure 6. **overview of our learnable Image-to-Hand-Params MDM.** Based on MDM [45], we extend the framework by incorporating environment information through an additional image condition branch. The model takes a noised hand pose sequence $\boldsymbol{p}_t^{1:L_p}$, t itself, a CLIP [37] text embedding $\boldsymbol{c}_{txt}$ of $\mathcal{T}$ and a CLIP image embedding $\boldsymbol{c}_{img}$ of $\boldsymbol{i}$ as input.

Motion Sequence Generation. MDM [45], utilizing a unique transformer encoder-only architecture, demonstrates remarkable performance in generating human motion sequences. Instead of predicting single-step noise, MDM directly generates clean motion sequences. Let $\boldsymbol{p}$ denotes motion sequences and $\mathcal{M}$ denotes MDM network. Here, $\boldsymbol{p}_0 \sim p_{data}(\boldsymbol{p})$ and $\boldsymbol{p}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The optimization objective 1 is modified as:

$$\min_{\theta_M} \mathbb{E}_{t,\boldsymbol{p},\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \|\boldsymbol{p}_0 - \mathcal{M}(\boldsymbol{p}_t, \mathcal{T}, t)\|_2^2, \tag{3}$$

where θ_M denotes the trainable parameters of MDM. After training, MDM generates the final clean motion sequence $\hat{\boldsymbol{p}}_0$ through a T-step denoising process. However, during each step, instead of directly generating $\boldsymbol{p}_{t-1}$ through single-step denoising, MDM firstly predicts the clean motion sequence $\hat{\boldsymbol{p}}_{0,t}$ from $\boldsymbol{p}_t$ with:

$$\hat{\boldsymbol{p}}_{0,t} = \mathcal{M}(\boldsymbol{p}_t, \mathcal{T}, t), \tag{4}$$

and then reintroduces noises to obtain $\boldsymbol{p}_{t-1}$. Through repeating this process for T times, MDM generates the final clean motion sequence $\hat{\boldsymbol{p}}_{0,0}$, denoted as $\hat{\boldsymbol{p}}_0$.

4.2. Stage I: A Coarse Action Planner

Our learnable coarse action planner is designed to generate a coarse HOI video $\hat{\boldsymbol{v}}_c \in \mathbb{R}^{L\times 3\times H\times W}$ conditioned on task description $\mathcal{T}$ and environment image $\boldsymbol{i}$. Specifically, we fine-tune DynamiCrafter [54] on TASTE-Rob to serve as the coarse action planner, denoted as $\mathcal{V}$.

Training. During fine-tuning, we adopt the training strategy similar to that of DynamiCrafter [54]. To leverage DynamiCrafter's robust temporal capabilities while adapting it to our specific HOI video generation, we only fine-tune its image context projector and spatial layers in its denoising U-Net. The training objective remains consistent with Equation 2, with trainable parameters θ_V representing parameters of the image context projector and spatial layers.

Hand-Object Interacting Inconsistency Problem. Our generated coarse HOI videos exhibit accurate task understanding, such as identifying the object to manipulate and determining its target position. However, as shown in $\boldsymbol{p}_c$ of Figure 5, the grasping poses exhibit temporal inconsistencies during manipulation, indicating a lack of motion coherence. Specifically, these inconsistencies refer to undesirable variations in grasping poses over time, manifesting as unnatural changes in hand posture when the relative position between the hands and the grasped object should ideally remain stable. As shown in $\boldsymbol{p}_c$ of Figure 5, the green hand pose demonstrates a pinching gesture, which is inconsistent with the grasping posture of the yellow hand pose and unsuitable for the target object being manipulated.

4.3. Stage II: Revising Hand Pose Sequences

To address HOI inconsistencies, we train an Image-to-Hand-Params MDM $\mathcal{M}$. This model refines hand pose sequences $\boldsymbol{p}_c \in \mathbb{R}^{L_p \times N_h \times 2}$ extracted from the coarse video $\hat{\boldsymbol{v}}_c$. Specifically, we define $\boldsymbol{p}_c$ as the normalized coordinates of hand key-point sequences, where L_p denotes the length of sequence and N_h is the number of hand key-points.

Training. Our learnable $\mathcal{M}$ is designed to take the task description $\mathcal{T}$ and environment image $\boldsymbol{i}$ as input to predict refined human hand pose sequences. To achieve this, we extend the original MDM [45] framework by incorporating environmental information through an additional image branch. As shown in Figure 6, similar to the text condition branch of [45], the added image branch integrates CLIP [37] class features of environment image $\boldsymbol{i}$. Consequently, the optimization objective 3 is updated as follows:

$$\min_{\theta_M} \mathbb{E}_{t,\boldsymbol{p},\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \|\boldsymbol{p}_0 - \mathcal{M}(\boldsymbol{p}_t, \mathcal{T}, \boldsymbol{i}, t)\|_2^2, \tag{5}$$

where θ_M denotes the parameters of $\mathcal{M}$ and ϵ denotes the sampled ground truth noise.

When we directly generate final clean pose sequence $\hat{\boldsymbol{p}}$ through a T-step denoising process, $\hat{\boldsymbol{p}}$ achieves physically plausible hand motions but exhibits limited spatial awareness. Conversely, $\boldsymbol{p}_c$ exhibits remarkable spatial awareness. To address this limitation, we refine $\boldsymbol{p}_c$ using $\mathcal{M}$ rather than generating from Gaussian noise. Specifically, we initialize the denoising process in Equation 4 of $\mathcal{M}$ with $\boldsymbol{p}_{0,N_{rv}}$, setting $\boldsymbol{p}_c$ as $\boldsymbol{p}_{0,N_{rv}}$. Through an N_{rv}-step denoising, we refine $\boldsymbol{p}_c$ to obtain the final clean hand pose sequence $\hat{\boldsymbol{p}}$ which satisfies both spatial accuracy and motion feasibility.

4.4. Stage III: Regenerating with Refined Pose

With the refined hand pose sequences, we generate the fine-grained HOI videos with the additional pose condition $\hat{\boldsymbol{p}}$. Inspired by ToonCrafter [55], we train a frame-independent pose encoder $\mathcal{S}$ to control the hand pose in the generated videos. We design $\mathcal{S}$ as a frame-wise adapter that adjusts the intermediate features of each frame independently, conditioned on $\hat{\boldsymbol{p}}$: $\boldsymbol{F}^i_{inject} = \mathcal{S}(\boldsymbol{s}^i, \boldsymbol{z}^i, t)$, where $\boldsymbol{s}^i$ is the visualized image sequences of $\hat{\boldsymbol{p}}$, shown as Figure 5, and $\boldsymbol{F}^i_{inject}$ is processed using a method similar to ControlNet [62].

Training. During training, we adopt a strategy similar to ToonCrafter [55], where all parameters of $\mathcal{V}$ are frozen, and only the parameters of $\mathcal{S}$, denoted as η, are trained with the following optimization objective:

$$\min_{\eta} \mathbb{E}_{t,\boldsymbol{s},\mathcal{E}(\boldsymbol{v}),\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \|\epsilon - \epsilon^{\mathcal{S}}_{\theta}(\boldsymbol{z}_t; \mathcal{T}, \boldsymbol{i}, \boldsymbol{s}, t, fps)\|_2^2, \tag{6}$$

where $\epsilon^{\mathcal{S}}_{\theta}$ denotes the combination of ϵ_θ and $\mathcal{S}$, and $\boldsymbol{s}$ represents the visualized images of hand pose sequences.

Finally, we consider our generated fine HOI videos $\hat{\boldsymbol{v}}$ as the video demonstrations for IL and enable robot manipulation using policy model of Im2Flow2Act [58]. As illustrated in Figure 10, we present the imitation learning results with our generated HOI videos as demonstrations, confirming their effectiveness for enabling robot manipulation.

5. Experiments

5.1. Implementation

During training, we train our model on the training set of TASTE-Rob dataset. **Stage I**: We fine-tune our coarse action planner based on DynamiCrafter for 30K steps, with a batch size of 16 and a learning rate of 5×10^{-5}. **Stage II**: We train our MDM for 100K steps, with a batch size of 64 and a learning rate of 1×10^{-4}. **Stage III**: We fine-tune the pose encoder based on SD for 30K steps, with a batch size of 32 and a learning rate of 5×10^{-5}. During inference, we generate videos with a 50-step denoising process and refine pose sequences with N_{rv} as 10 (See more details in Appendix 12).

5.2. Evaluation

Datasets. We evaluate on two datasets: TASTE-Rob-Test and Mujoco simulation dataset. **TASTE-Rob-Test**: We split TASTE-Rob dataset into training and test sets, with the test set denoted as TASTE-Rob-Test. For fair evaluation, we select 2% of samples from each action category as test samples. **Mujoco simulation dataset**: we randomly select 50 tasks from robot simulation platform provided by Im2Flow2Act [58] for robot manipulation.

In addtion, as mentioned in Section 3 and Table 2, other ego-centric video datasets feature different settings from our task, making them unsuitable as evaluation datasets.

Metrics. We evaluate the performance on three aspects: video generation, grasping pose consistency and robot manipulation. **Video Generation**: To evaluate the quality of

generated videos in both the spatial and temporal domains, we report Fréchet Video Distance (FVD) [46], Kernel Video Distance (KVD) [46] and Perceptual Input Conformity (PIC) [54] as used in DynamiCrafter [54]. **Grasping Pose Consistency**: We evaluate with three metrics - **Grasping Pose Variance** (GPV, computed by $\frac{1}{L^r}\sum_{i=1}^{N}(\theta_i^p - \bar{\theta^p})^2$), **Grasp Type Classification Error** (GTCE, calculating the classification error of grasping types in generated videos) and **Hand Movement Direction Accuracy** (HMDA, Measuring the directional consistency of hand movements and their alignment with the intended task trajectory), and details of these metrics are illustrated in Appendix 10. **Robot Manipulation** : We utilize videos generated on the Mujoco simulation dataset as demonstrations for IL, and deploy the policy model of Im2Flow2Act [58] for robot manipulation. The performance is evaluated by comparing the success rates.

Method	Generation Quality		
	KVD↓	FVD↓	PIC↑
consistI2V	0.24	65.77	0.85
DynamiCrafter	0.21	62.36	0.79
Open-Sora Plan	0.18	50.19	0.84
CogVideoX	0.16	48.72	0.85
TASTE-Rob	**0.03**	**9.43**	**0.90**

Table 3. **Quantitative comparison between TASTE-Rob and baselines.** All video generation metrics prove the better generation performance of TASTE-Rob.

5.3. Baselines and Comparisons.

We select four existing powerful I2V diffusion models - DynamiCrafter [54], consistI2V [38], Open-Sora Plan [26] and CogVideoX [60], as baselines, and conduct comparative experiments among these baselines and our method. We provide qualitative comparisons of video generation performance on TASTE-Rob-Test and wild environment in Figure 7, and quantitative comparisons on TASTE-Rob-Test in Table 3, which demonstrates superior video quality and better generalization capability of our method. According to above experiments, all existing powerful general VDMs fail to achieve manipulation tasks well, making them unsuitable for generating HOI video demonstrations. Given that the other two evaluation aspects focus on measuring the fine-grained details of generated videos, we omit further comparisons with these baseline methods.

5.4. Ablation Study

TASTE-Rob Dataset. We evaluate the performance improvements achieved by using our TASTE-Rob dataset. To be fair, we compare video generation performance between our coarse action planner (Coarse-TASTE-Rob) and Ego4D-Gen, a comparison baseline fine-tuned from DynamiCrafter [54] on 83,647 ego-centric HOI videos from Ego4D [16]. The only difference between them is the training datasets they use. As shown in Figure 8 and Table 4, our coarse action planner achieves superior performance.

Method	Generation Quality		
	KVD↓	FVD↓	PIC↑
Ego4D-Gen	0.18	52.17	0.77
Coarse-TASTE-Rob	**0.04**	**10.85**	**0.88**

Table 4. **Quantitative comparison between Ego4D-Gen and Coarse-TASTE-Rob.** All video generation metrics prove the better generation performance of Coarse-TASTE-Rob.

Metric	Coarse-TASTE-Rob	TASTE-Rob
KVD↓	0.04	**0.03**
FVD↓	10.85	**9.43**
PIC↑	0.88	**0.90**
GPV↓	0.28	**0.24**
GTCE↓	67.8%	**9.7%**
HMDA↓	26.4°	**11.3°**
success rate↑	84%	**96%**

Table 5. **Quantitative comparison between TASTE-Rob and Coarse-TASTE-Rob.** All metrics prove the better generation performance, grasping pose consistency and more precise robot manipulation of TASTE-Rob.

Pose-Refinement Pipeline. We compare Coarse-TASTE-Rob and TASTE-Rob across all three evaluation aspects, as shown in Table 5. **At a coarse level**, both the video generation metrics in Table 5 and the visual comparisons in Figure 9 demonstrate that our proposed pose-refinement pipeline achieves better HOI video generation. **At a fine level**, both the grasping pose consistency metrics in Table 5 and Figure 9 further validate that our proposed pose-refinement pipeline effectively addresses the pose inconsistency issue and enables more precise robot manipulation. Besides, as shown in Figure 10 and success rate metrics in Table 5, our proposed pose-refinement pipeline helps the policy model achieve more accurate target placement, leading to improved success rates. In addition, we only do this ablation on our dataset, because videos generated by fine-tuning on existing HOI datasets (e.g., Ego4D [16]) are of low quality so that it's difficult to extract reasonable poses for further refinement.

Denoising Steps of Pose Refinement. We also discuss the selection of N_{rv} during pose refinement. When $N_{rv} = 0$ (no refinement), hand pose sequences exhibit inconsistency issues, while $N_{rv} = 50$ (generating pose sequences from Gaussian noise) results in limited spatial awareness.

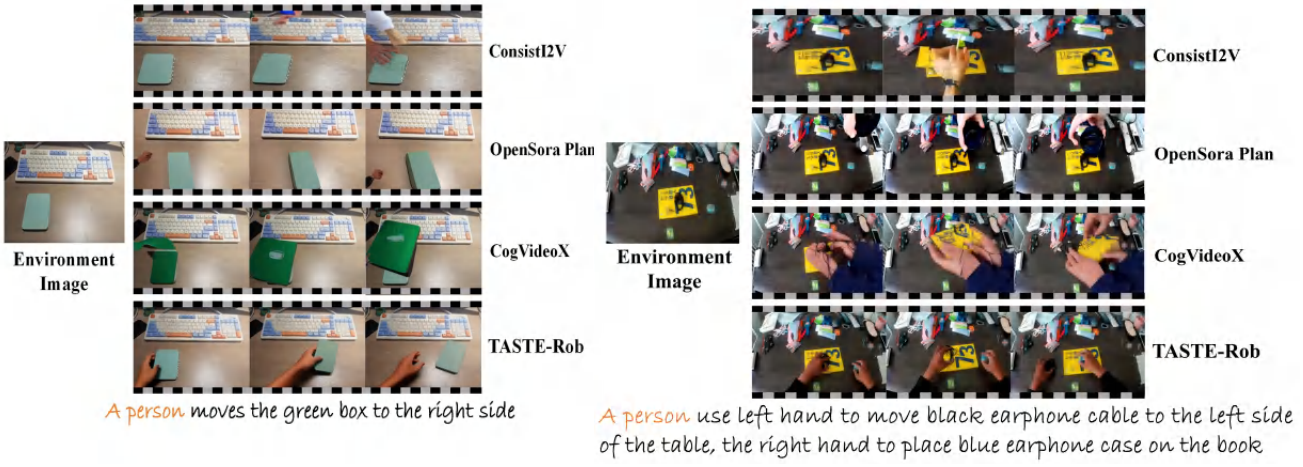

Figure 7. **Comparisons of video generation performance.** TASTE-Rob generates videos with high-quality and accurate action, while other existing powerful general VDMs fail.

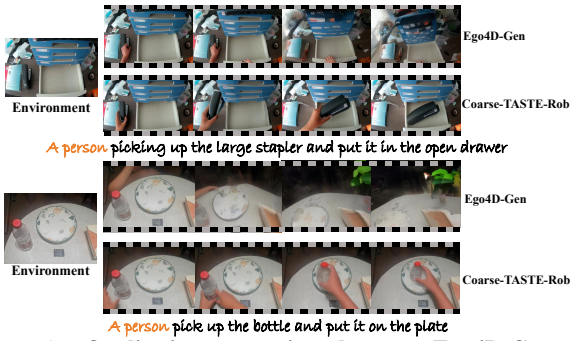

Figure 8. **Qualitative comparison between Ego4D-Gen and Coarse-TASTE-Rob.** Our TASTE-Rob dataset significantly improves the generalization of video generation models in generating HOI videos with accurate actions.

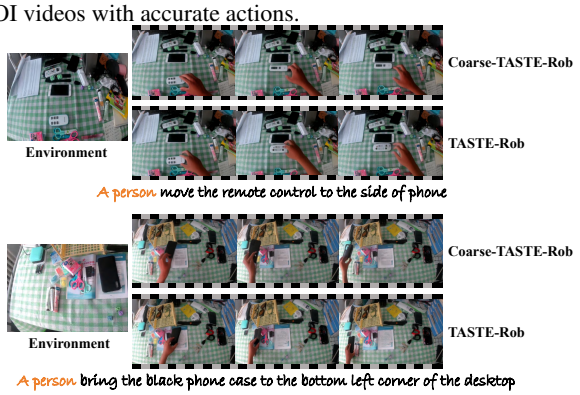

Figure 9. **Qualitative comparison between TASTE-Rob and Coarse-TASTE-Rob.** With our proposed pose-refinement pipeline, we can generate HOI videos with more accurate and feasible grasping poses.

As N_{rv} increases, we observe improved pose consistency but degraded spatial awareness. As shown in Table 6, when N_{rv} increases, video generation quality first improves and then deteriorates. Based on this trade-off, we finally select the value of N_{rv} as 10 (see more details in Appendix 11).

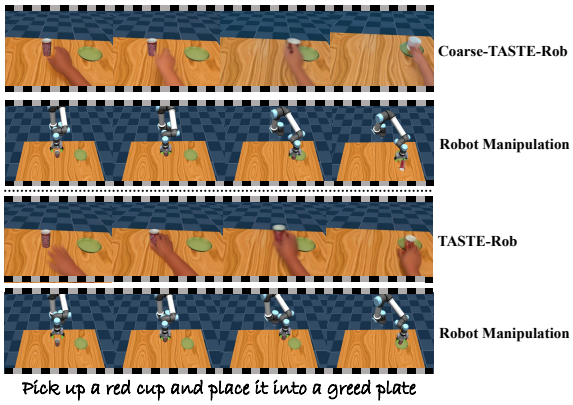

Figure 10. **Comparison of Robot Manipulation Performance.** TASTE-Rob generates higher-quality HOI videos, leading to more precise robot manipulation.

6. Discussion and Conclusion

Summary. In this work, we collect a diverse, large-scale ego-centric HOI dataset and develop a three-stage pose-refinement pipeline to generate high-quality task-oriented HOI videos. We successfully enhance HOI video generation capabilities, thereby improving robot manipulation generalization to novel scenes through imitation learning.

Limitations. Our approach achieves accurate manipulation on unseen scenes and enhances the grasping poses in generated videos. However, the generation quality degrades when objects undergo significant transformations, such as rotation and opening. In our future work, we plan to address this limitation.

7. Acknowledgments

The work was supported in part by the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen-HK S&T Cooperation Zone, Guangdong Provincial Outstanding Youth Fund(No. 2023B1515020055), NSFC-61931024, and Shenzhen Science and Technology Program No. JCYJ20220530143604010. It is also partly supported by the National Key R&D Program of China with grant No. 2018YFB1800800, by Shenzhen Outstanding Talents Training Fund 202002, by Guangdong Research Projects No. 2017ZT07X152 and No. 2019CX01X104,by Key Area R&D Program of Guangdong Province (Grant No. 2018B030338001), by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001), and by Shenzhen Key Laboratory of Big Data and Artificial Intelligence (Grant No. ZDSYS201707251409055).

References

[1] Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. 2022. 1

[2] Homanga Bharadhwaj, Abhinav Gupta, Vikash Kumar, and Shubham Tulsiani. Towards generalizable zero-shot manipulation via translating human interaction plans, 2023. 1

[3] Homanga Bharadhwaj, Abhinav Gupta, and Shubham Tulsiani. Visual affordance prediction for guiding robot exploration. *IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 1

[4] Homanga Bharadhwaj, Jay Vakil, Mohit Sharma, Abhinav Gupta, Shubham Tulsiani, and Vikash Kumar. Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking, 2023. 1

[5] Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation, 2024. 1, 2

[6] Homanga Bharadhwaj, Roozbeh Mottaghi, Abhinav Gupta, and Shubham Tulsiani. Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation. In *European Conference on Computer Vision (ECCV)*, 2024.

[7] Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pretrained image-editing diffusion models, 2023. 1, 2

[8] Anthony Brohan, Noah Brown, Justice Carbajal, and *et al*. Rt-1: Robotics transformer for real-world control at scale. In *arXiv preprint arXiv:2212.06817*, 2022. 1, 2

[9] Elliot Chane-Sane, Cordelia Schmid, and Ivan Laptev. Learning video-conditioned policies for unseen manipulation tasks. In *ICRA*, 2023. 1, 2

[10] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023. 1, 2

[11] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. 2, 4

[12] Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation, 2023. 1

[13] Alircza Fathi, Jessica K. Hodgins, and James M. Rehg. Social interactions: A first-person perspective. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1226–1233, 2012. 2, 4

[14] Thomas Feix, Javier Romero, Heinz-Bodo Schmiedmayer, Aaron M. Dollar, and Danica Kragic. The grasp taxonomy of human grasp types. *IEEE Transactions on Human-Machine Systems*, 46(1):66–77, 2016. 4

[15] Pete Florence, Corey Lynch, Andy Zeng, Oscar Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. Implicit behavioral cloning. *Conference on Robot Learning (CoRL)*, 2021. 1, 2

[16] Kristen Grauman, Andrew Westbury, and *et al*. Ego4d: Around the world in 3,000 hours of egocentric video, 2022. 1, 2, 3, 4, 7

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 5

[18] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022. 2

[19] Vincent Tao Hu, Wenzhe Yin, Pingchuan Ma, Yunlu Chen, Basura Fernando, Yuki M Asano, Efstratios Gavves, Pascal Mettes, Bjorn Ommer, and Cees G. M. Snoek. Motion flow matching for human motion synthesis and editing, 2023. 2

[20] Vidhi Jain, Maria Attarian, Nikhil J Joshi, Ayzaan Wahid, Danny Driess, Quan Vuong, Pannag R Sanketi, Pierre Sermanet, Stefan Welker, Christine Chan, Igor Gilitschenski, Yonatan Bisk, and Debidatta Dwibedi. Vid2robot: End-to-end video-conditioned policy learning with cross-attention transformers, 2024. 1, 2

[21] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. BC-z: Zero-shot task generalization with robotic imitation learning. In *5th Annual Conference on Robot Learning*, 2021. 1

[22] Ananth Jonnavittula, Sagar Parekh, and Dylan P. Losey. View: Visual imitation learning with waypoints, 2024. 1, 2

[23] Siddharth Karamcheti, Suraj Nair, Annie S. Chen, Thomas Kollar, Chelsea Finn, Dorsa Sadigh, and Percy Liang. Language-driven representation learning for robotics. In *Robotics: Science and Systems (RSS)*, 2023. 2

[24] Alexander Khazatsky, Karl Pertsch, and *et al*. Droid: A large-scale in-the-wild robot manipulation dataset, 2024. 1

[25] Po-Chen Ko, Jiayuan Mao, Yilun Du, Shao-Hua Sun, and Joshua B. Tenenbaum. Learning to act from actionless videos through dense correspondences, 2023. 1

[26] PKU-Yuan Lab and Tuzhan AI etc. Open-sora-plan, 2024. 2, 7

[27] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1346–1353, 2012. 2, 4

[28] Yin Li, Miao Liu, and James M. Rehg. In the eye of the beholder: Gaze and actions in first person video, 2020. 2, 4

[29] 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 (CVPR)*, pages 21013–21022, 2022. 4

[30] Zheng Lu and Kristen Grauman. Story-driven summarization for egocentric video. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2714–2721, 2013. 2, 4

[31] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for building perception pipelines, 2019. 4, 5

[32] Yecheng Jason Ma, William Liang, Vaidehi Som, Vikash Kumar, Amy Zhang, Osbert Bastani, and Dinesh Jayaraman. Liv: Language-image representations and rewards for robotic control, 2023. 2

[33] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models, 2023. 2

[34] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024. 4, 2, 3

[35] 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, 2012. 2, 4

[36] Sigal Raab, Inbal Leibovitch, Guy Tevet, Moab Arar, Amit H Bermano, and Daniel Cohen-Or. Single motion diffusion. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. 2

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 5, 6

[38] Weiming Ren, Huan Yang, Ge Zhang, Cong Wei, Xinrun Du, Wenhao Huang, and Wenhu Chen. Consisti2v: Enhancing visual consistency for image-to-video generation, 2024. 2, 7

[39] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6):1–17, 2017. 2

[40] Dandan Shan, Jiaqi Geng, Michelle Shu, and David Fouhey. Understanding human hands in contact at internet scale. In *CVPR*, 2020. 3

[41] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning, 2023. 2

[42] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Proceedings of the 6th Conference on Robot Learning (CoRL)*, 2022. 1, 2

[43] 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, 2018. 2, 4

[44] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015. 5

[45] 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. 2, 4, 5, 6

[46] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. In *DGS@ICLR*, 2019. 7

[47] Boyang Wang, Nikhil Sridhar, Chao Feng, Mark Van der Merwe, Adam Fishman, Nima Fazeli, and Jeong Joon Park. This&that: Language-gesture controlled video generation for robot planning, 2024. 1

[48] Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play, 2023. 1, 2

[49] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability, 2023. 4

[50] Yin Wang, Zhiying Leng, Frederick W. B. Li, Shun-Cheng Wu, and Xiaohui Liang. Fg-t2m: Fine-grained text-driven human motion generation via diffusion model. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21978–21987, 2023. 2

[51] Zhiqiang Wang, Hao Zheng, Yunshuang Nie, Wenjun Xu, Qingwei Wang, Hua Ye, Zhe Li, Kaidong Zhang, Xuewen Cheng, Wanxi Dong, Chang Cai, Liang Lin, Feng Zheng, and Xiaodan Liang. All robots in one: A new standard and unified dataset for versatile, general-purpose embodied agents, 2024. 1

[52] Chuan Wen, Xingyu Lin, John So, Kai Chen, Qi Dou, Yang Gao, and Pieter Abbeel. Any-point trajectory modeling for policy learning, 2024. 2

[53] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. In *International Conference on Learning Representations*, 2024. 2

[54] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors, 2023. 2, 5, 7, 4

[55] Jinbo Xing, Hanyuan Liu, Menghan Xia, Yong Zhang, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Tooncrafter: Generative cartoon interpolation. *ACM Transactions on Graphics (TOG)*, 43(6):1–11, 2024. 2, 6, 5

[56] Haoyu Xiong, Quanzhou Li, Yun-Chun Chen, Homanga Bharadhwaj, Samarth Sinha, and Animesh Garg. Learning by watching: Physical imitation of manipulation skills from human videos, 2021. 1, 2

[57] Mengda Xu, Zhenjia Xu, Cheng Chi, Manuela Veloso, and Shuran Song. XSkill: Cross embodiment skill discovery. In *7th Annual Conference on Robot Learning*, 2023. 1, 2

[58] Mengda Xu, Zhenjia Xu, Yinghao Xu, Cheng Chi, Gordon Wetzstein, Manuela Veloso, and Shuran Song. Flow as the cross-domain manipulation interface. In *8th Annual Conference on Robot Learning*, 2024. 1, 2, 6, 7

[59] Zhao Yang, Bing Su, and Ji-Rong Wen. Synthesizing long-term human motions with diffusion models via coherent sampling, 2023. 2

[60] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihan Wang, Yean Cheng, Ting Liu, Bin Xu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer, 2024. 2, 7, 3, 4

[61] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. 2

[62] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 2, 6

[63] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models, 2023. 4

[64] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware, 2023. 1, 2

[65] Yifeng Zhu, Arisrei Lim, Peter Stone, and Yuke Zhu. Vision-based manipulation from single human video with open-world object graphs, 2024. 1