Grounding Generated Video Plans in Simulation Towards Versatile Dexterous Controllers

Tianyue Wu[2,3,1,*,‡], Boyuan An[2,1,*], Shuqi Zhao[1], Heyu Guo[2], Wanli Xing[2], Yi Ma[3,1], Kaifeng Zhang[2], Ruihai Wu[1,†], and Masayoshi Tomizuka[1,†]

[1]University of California, Berkeley [2]Sharpa [3]The University of Hong Kong
[*]Co-first authors [†]Co-advisors [‡]Corresponding author

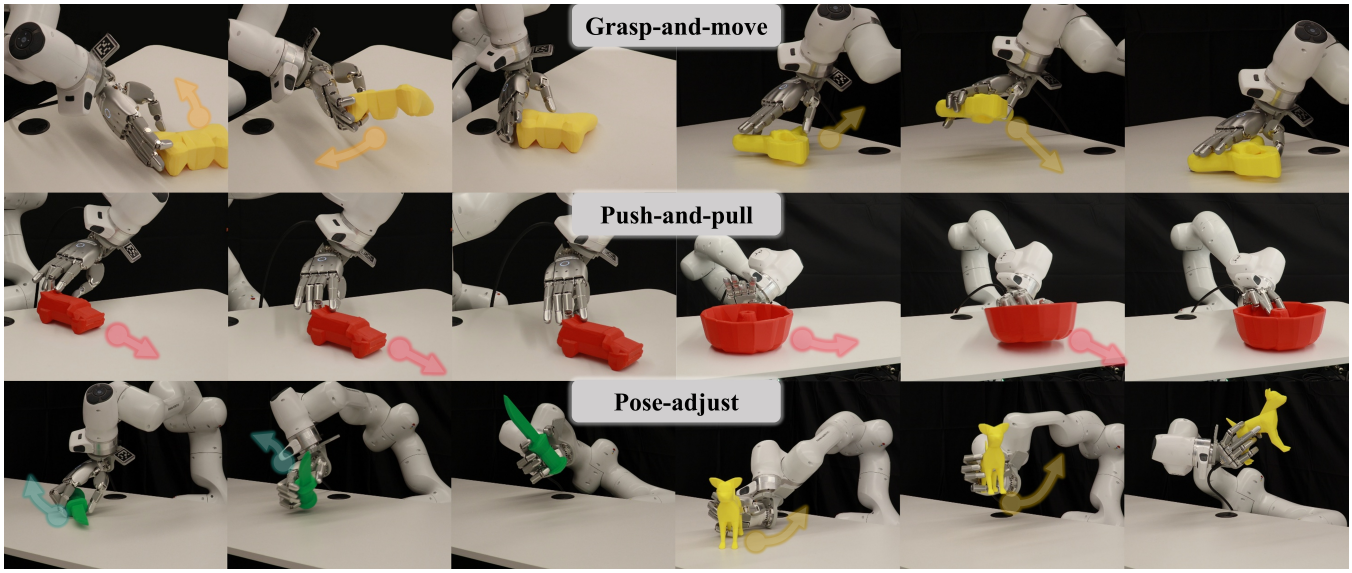

Fig. 1. Real-world execution of generated video references. A unified tracking controller performs grasp-and-move, push-and-pull, and pose-adjust behaviors across diverse objects.

***Abstract*— Generated hand–object interaction (HOI) videos provide a controllable way to propose manipulation motions. Simulation-based HOI tracking can translate such kinematic references into feasible low-level control, but its scalability is limited by the lack of reliable reference motions. We therefore combine generated videos with simulation-based HOI grounding: during training, generated videos provide diverse motion references for learning a multi-object, multi-trajectory HOI tracker, and at deployment, the video model produces motion plans that are executed by the learned tracker. In particular, we propose a method that enables scalable reference generation by HOI reconstruction with minimal manual intervention and successfully grounds more than 1,500 generated videos in simulation, achieving success rates over 25 percentage points higher than those of baselines during simulation-based training. In real-world closed-loop experiments, it achieves diverse grasps, including functional grasps, non-prehensile manipulation, and post-grasp object-pose tracking. Videos and code are available at this URL.**

I. INTRODUCTION

Human children and young animals learn many manipulation skills by watching others. Although they cannot observe the demonstrator's motor commands, they can see both how a similar body moves and how those movements affect the surrounding world. Together, these visual cues provide dense, indirect supervision for learning purposeful behavior [1], [2]. These observations motivate a promising methodology in robotics: dexterous manipulation skill acquisition from visual demonstrations [3], [4].

In-the-wild videos contain hand behaviors at a scale and diversity beyond current robot demonstration datasets, but robots cannot interpret them as humans do. The intended interactions are entangled with irrelevant human motion and uncontrolled variation in viewpoint, appearance, occlusion, and scene context, so extracting useful references still requires extensive human filtering and annotation [5]. Prior work obtains cleaner supervision by recording human manipulation with standardized hardware and prescribed task protocols [6], but collection, annotation, and validation remain costly. Moreover, videos do not directly reveal 3-D hand–object states, contacts, forces, or actions executable by a particular robot embodiment. Moving toward versatile dexterous control through these visual demonstrations therefore requires both a more scalable and controllable source of diverse references and a video retargeter that can physically ground them.

Accordingly, we study simulation-trained multi-object, multi-trajectory hand–object tracking as a scalable task [7], with video generation models [8] supplying its training references. Advanced video foundation models provide reference

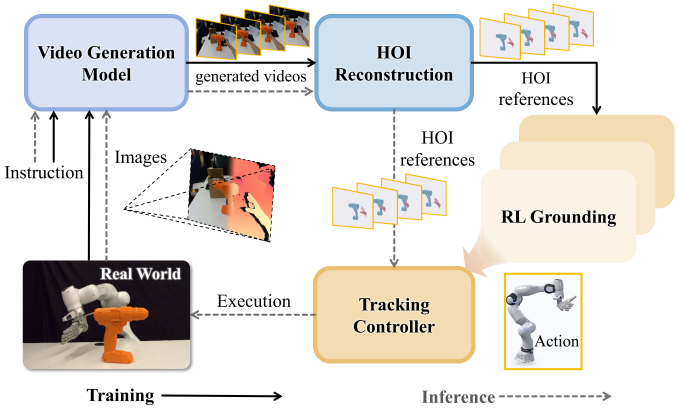

Fig. 2. Overview of GALATEA.

scale by exposing priors learned from Internet-scale video corpora through controlled synthesis: image and language conditioning can specify the object, intended behavior, and scene while suppressing irrelevant variation, producing diverse and controllable manipulation references with less manual curation compared to in-the-wild videos. Modern simulators [9] provide a complementary path to scale by enabling parallel interaction rollouts, which supply dense supervision for mapping kinematic hand–object references to dynamically feasible low-level actions. The resulting general low-level controller can serve as a physics-aware execution layer for high-level robot foundation models [7]. However, existing video-to-manipulation methods typically reconstruct or optimize only a limited set of references [10], [11], while learning a versatile HOI tracker from relatively large, noisy, video-derived reference corpora remains underexplored.

Here, we present GALATEA[1] to Ground generAted-video pLans At scale with simulation-trained Tracking for Executable Actions. During training, generated videos provide diverse motion references for learning a multi-object, multi-trajectory tracker; during deployment, prompted video plans are physically executed by the learned tracker, as shown in Fig. 2. GALATEA consists of two stages. First, a reconstruction pipeline combines foundation-model perception, stereo initialization, and joint hand–object optimization to recover HOI trajectories from generated videos. Second, a sim-to-real RL recipe combines a task-agnostic tracking objective, reference augmentation, and the Split and Aggregate Policy Gradients (SAPG) [12] optimizer, outperforming baselines by more than 25 percentage points of success rates. Finally, GALATEA reconstructs about 2,000 usable HOI trajectories from 2,500 generated clips and more than 1,500 trajectories are successfully grounded in simulation by expert controllers, which are distilled into a unified policy.

Our contributions are summarized as follows:

- GALATEA, a system that bridges generated HOI video plans and physical execution through HOI reconstruction and a motion tracking controller;
- a practical HOI reference extraction method from generated videos combining advanced perception model, stereo depth, and joint hand–object optimization;
- an improved tracking-style RL recipe with design choices in RL formulation and the optimizer to support learning from multiple noisy, video-derived HOI references;
- closed-loop real-world execution of both trained and unseen video plans, which achieves diverse grasps, pushing/pulling, and post-grasp object pose adjustment (Fig. 1).

[1]Named after Galatea, the statue brought to life in later retellings of the Pygmalion myth, echoing our goal of turning imagined videos into reality.

II. RELATED WORK

A. Generative Video Models for Manipulation

Generative video priors enter manipulation policies through several approaches. Some approaches generate visual futures and translate them into robot actions with an inverse-dynamics model or a video-conditioned policy [13], [14], whereas others jointly model visual futures and actions [15], [16]. Their action components typically require embodiment-specific robot demonstrations for training. Methods with an explicit intermediate interface instead convert generated videos into structured motion references for downstream control [17]. LVP [18] retargets generated wrist and hand kinematics, but does not track object motion or contact during open-loop execution. Dex4D [19] conditions a closed-loop, simulation-trained policy on object-centric 3-D point tracks, but leaves the desired interaction strategy under-specified, limiting control over functional grasps. We instead reconstruct and physically ground both finger-level and object trajectories, preserving both the demonstrated agent motion and its intended effect on the object. Related work also uses generative video models as motion planners or data generators for humanoid control [20], [21], [22], rather than for contact-rich dexterous manipulation. Concurrent works on dexterous manipulation systems that are grounded in the ”System 1/System 2” framework [23] also employ single-task video imitation policies [24] or Vision-Language-Action (VLA) models [25] as motion planners. By integrating a video-based model with fundamentally superior generalization capabilities, we retain the possibility of achieving zero-shot manipulation.

B. Retargeting from Kinematic Demonstrations for Dexterous Control

Kinematic demonstrations provide dense supervision over desired hand and object motion, while simulation can recover the contact-feasible actions absent from such references. Existing approaches can be grouped by the source of their kinematic supervision. One line retaregts from high-quality motion capture or otherwise curated human–object trajectories [26], [27], [28], [29], [30], [31], [32]. A second line reconstructs references from one or a small number of human videos [33], [10], [34], [35], [11], [36], where the trajectories can be noisier than those in the first line and are hard to solve using methods like open-loop motion planning. A third line, which our work belongs to, reduces dependence on manually captured trajectories through synthetic or simulation-expanded trajectories [37], [38], [24], which are more diverse and scalable but introduce additional noise

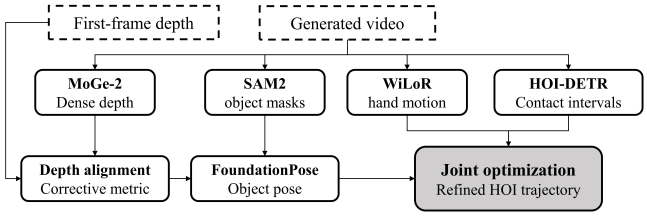

Fig. 3. HOI reconstruction pipeline. First-frame depth metric-aligns generated-video depth. Object masks, initial hand and object motion, and contact intervals then guide joint optimization to produce final estimates.

or artifacts. Across these previous works, only a limited subset using RL to retarget from a relatively large set of HOI references [30], [38], some of which is limited to simulation dynamics [37]. In contrast, we demonstrate that a unified controller can be learned from a relatively large-scale and noisy corpus of generated-video references, and empirically establish the superiority of our RL training recipe over baseline methods.

III. METHOD

Our method has two stages. The first generates videos and reconstructs their 3-D hand–object trajectories. The second uses tracking-style RL and policy distillation in simulation to learn a unified controller from these trajectories.

A. Video Generation and Reconstruction

1) Video Generation from Collected Images: Minimizing manual effort requires both video generation and reconstruction to succeed reliably. We therefore use the state-of-the-art proprietary video model Seedance 2.0 [8], which produces substantially more plausible manipulation videos than current open-source alternatives in our tests. Because simulation-rendered conditioning images often lead to physically inconsistent interactions, we instead capture a real first-frame RGB image and condition the model on this image and a language instruction. We consider three manipulation types. *Grasp-and-move* grasps an object from the table, moves it through space with little change in orientation, and may place it back on the table. *Push-and-pull* moves the object on the table without lifting. *Pose-adjust* grasps the object and substantially reorients it in midair. A stereo camera simultaneously captures first-frame depth, which provides a metric anchor for subsequent reconstruction.

2) HOI Reconstruction: We use H and O to denote the hand and object, respectively, and $X \in \{H, O\}$ to index either entity. Given a generated video $\{I_t\}_{t=1}^{T}$, calibrated camera intrinsics $\mathbf{K}$, a metric object mesh $\mathcal{M}^O$, and first-frame stereo depth D_1^{st}, we reconstruct an aligned MANO [39] hand mesh and 6-DoF object trajectory by estimating both motions independently and then jointly refining them, inspired by prior joint hand–object reconstruction methods [40], [41], which is crucial for the success of bootstrap learning. Fig. 3 summarizes the pipeline.

Metric Depth Estimation. MoGe-2 [42] predicts a dense depth map $\widetilde{D}_t$, whose global scale and offset may drift across frames. SAM2 [43] provides object masks M_t^O and moving-foreground masks M_t^F, with M_t^F covering the hand, forearm, and object. Under the fixed-camera and static-background assumption, Ω_t^{bg} denotes pixels outside both M_1^F and M_t^F, which observe the same background in the first and current frames. Following [22], we use trimmed least squares over Ω_t^{bg} to fit $s_t\widetilde{D}_t + b_t$ to the first-frame stereo depth D_1^{st}, and set $D_t(\mathbf{p}) = [s_t\widetilde{D}_t(\mathbf{p}) + b_t]_+$. This anchors each frame to a corrective metric scale and offset. D_t is then passed to object tracking. RANSAC [44] on D_1^{st} also defines a gravity-aligned support plane, i.e., the table.

Initial Motion Estimation. FoundationPose [45] estimates initial object poses $\widehat{\mathbf{T}}_t^O$ from $(I_t, D_t, M_t^O, \mathcal{M}^O, \mathbf{K})$; low mask coverage triggers a color-adapted retry. WiLoR [46] estimates MANO vertices $\widehat{\mathbf{V}}_t^H$ and 2D keypoints $\mathbf{u}_t$.

Joint Hand–Object Optimization. Starting from the initial estimates, we optimize per-frame rigid corrections $\mathbf{T}_t^O = \Delta\mathbf{T}_t^O\widehat{\mathbf{T}}_t^O$ and $\mathbf{V}_t^H = \Delta\mathbf{T}_t^H\widehat{\mathbf{V}}_t^H$, while retaining finger articulation. Removing object and forearm regions from M_t^F gives the observed hand mask M_t^H. For $X \in \{O, H\}$, a differentiable renderer projects the current mesh into a filled 2-D occupancy mask S_t^X, which we call its silhouette. We match it to the observed mask M_t^X using $\mathcal{L}_{\mathrm{sil}}^X = \sum_t \|W_t^X \odot (S_t^X - M_t^X)\|_2^2$, where W_t^X excludes pixels occluded by the other component. Given metric mesh dimensions and camera intrinsics, the mask position, shape, and size constrain lateral translation, rotation, and depth. HOI-DETR [47] detects contact frames $\mathcal{T}_c$, which gate the contact loss $\mathcal{L}_{\mathrm{con}}$. The full objective is

$$\begin{aligned}\mathcal{L} = {} & \lambda_{\mathrm{FP}}\mathcal{L}_{\mathrm{FP}} + \lambda_{\mathrm{kp}}\mathcal{L}_{\mathrm{kp}} + \lambda_{\mathrm{sil}}^O\mathcal{L}_{\mathrm{sil}}^O \\ & + \lambda_{\mathrm{sil}}^H\mathcal{L}_{\mathrm{sil}}^H + \lambda_{\mathrm{con}}\mathcal{L}_{\mathrm{con}} + \lambda_{\mathrm{temp}}\mathcal{L}_{\mathrm{temp}}.\end{aligned} \tag{1}$$

Let $\pi_{\mathbf{K}}$ denote perspective projection, $\mathbf{V}^O$ the object vertices, $\mathbf{J}_t^H$ the MANO joints, $\mathbf{c}_t^O$ the object translation, $\mathbf{c}_t^H$ the hand-mesh centroid, and $\mathcal{C}_t$ the contact-labeled hand-segment vertices. The remaining losses are

$$\begin{aligned}\mathcal{L}_{\mathrm{FP}} &= \sum_t \left\| \pi_{\mathbf{K}}(\mathbf{T}_t^O\mathbf{V}^O) - \pi_{\mathbf{K}}(\widehat{\mathbf{T}}_t^O\mathbf{V}^O) \right\|_1, \\ \mathcal{L}_{\mathrm{kp}} &= \sum_t \left\| \pi_{\mathbf{K}}(\mathbf{J}_t^H) - \mathbf{u}_t \right\|_1, \\ \mathcal{L}_{\mathrm{con}} &= \sum_{t\in\mathcal{T}_c}\sum_{\mathbf{v}\in\mathcal{C}_t} d^2(\mathbf{v}, \mathbf{T}_t^O\mathcal{M}^O), \\ \mathcal{L}_{\mathrm{temp}} &= \sum_{X\in\{O,H\}}\sum_t \left(\|\Delta\mathbf{c}_t^X\|_1 + \beta\|\Delta^2\mathbf{c}_t^X\|_1 \right).\end{aligned} \tag{2}$$

Here $d(\cdot,\cdot)$ is point-to-mesh distance and Δ is the temporal difference operator. The first two terms preserve the FoundationPose object projection and WiLoR keypoint reprojection. The contact term draws the hand to the object surface, and the temporal term penalizes velocity and acceleration. We assign a high weight to $\mathcal{L}_{\mathrm{FP}}$ because the object estimates are relatively reliable and smooth in our controlled setting.

B. Grounding Generated Videos in Simulation with HOI Tracking

We use a Sharpa Wave Hand [48] mounted on a Franka Research 3 arm [49], simulated with PhysX in Isaac Gym [9].

1) Reference Augmentation and Preprocessing: We first augment each source trajectory with 5 sampled variations in hand approach and post-contact object motion:

$$\begin{aligned}
\widetilde{\mathbf{T}}_t^X &= \mathbf{G}\mathbf{C}_t^X\mathbf{T}_t^X, \quad X \in \{H, O\}, \\
\widetilde{\mathbf{J}}_t^H &= (\mathbf{G}\mathbf{C}_t^H)\cdot\mathbf{J}_t^H, \\
(\mathbf{C}_t^H, \mathbf{C}_t^O) &= \begin{cases} (\mathrm{Exp}(\alpha_t\boldsymbol{\xi}_0), \mathbf{I}), & t < t_c, \\ (\mathrm{Exp}(\beta_t\boldsymbol{\xi}_e), \mathrm{Exp}(\beta_t\boldsymbol{\xi}_e)), & t \geq t_c. \end{cases}
\end{aligned} \tag{3}$$

Here, t_c and T denote first contact and the final frame. $\mathbf{T}_t^X$ and $\mathbf{J}_t^H$ are the source poses and MANO points, and tildes denote augmented quantities. $\mathbf{C}_t^X$ is the perturbation and $\mathbf{G}$ a trajectory-level yaw transform. The exponential map maps sampled twists to $SE(3)$. Before contact, $\alpha_t = 1 - t/t_c$ maps $[0, t_c]$ from 1 to 0. Afterward, $\beta_t = (t - t_c)/(T - t_c)$ maps $[t_c, T]$ from 0 to 1. We set both perturbations to 30% of the source clearance and motion range. Sharing the post-contact transform preserves the hand–object relative pose during contact. Each augmented candidate is then screened for ease of execution. We solve its FR3 trajectory using inverse kinematics (IK) and slow the whole trajectory with interpolation according to defined joint-velocity limits.

2) RL Formulation and Training Recipe: **Observation and control.** We train an asymmetric actor–critic RL to learn the tracking policies. We design the actor input as

$$\mathbf{o}_t^\pi = [\,\mathbf{q}_t, \mathbf{a}_{t-1}, \mathbf{x}_t^w, \mathbf{x}_t^{\mathrm{tip}}, \widetilde{\mathbf{x}}_t^o, \mathbf{e}_t^{\mathrm{HOI}}, \mathbf{d}_t^{\mathrm{ref}}, \phi(\mathcal{M}^o)\,], \tag{4}$$

where $\mathbf{a}_{t-1}$ is the previous raw action, wrist quantities are expressed in the FR3 base frame, fingertips and the noisy object pose are expressed relative to the wrist, and $\mathbf{e}_t^{\mathrm{HOI}}$ contains current wrist, MANO-keypoint, and object-pose errors. The remaining terms are the five reference fingertip–surface distances and a BPS object-shape encoding. Inspired by [27], we redundantly encode hand–object relative geometry such as HOI errors in the observation space. We find that this simplifies exploration. The actor receives no joint velocity, force, mass, center of mass, or retargeted hand-joint target. The critic additionally receives privileged simulator state including clean object pose and joint velocity.

The policy is represented as an MLP that outputs $\mathbf{a}_t = [\mathbf{a}_t^A, \mathbf{a}_t^H] \in [-1, 1]^{29}$ (seven arm and 22 hand dimensions). Arm actions are joint deltas about the measured state, $\bar{\mathbf{q}}_t^A = \mathbf{q}_t^A + (0.03\,\mathrm{rad})\mathbf{a}_t^A$. Hand actions are mapped through the joint limits to absolute targets $\bar{\mathbf{q}}_t^H$. Both targets are smoothed by an exponential moving average (EMA), $\mathbf{q}_t^{X,\mathrm{PD}} = \alpha_X\bar{\mathbf{q}}_t^X + (1 - \alpha_X)\mathbf{q}_{t-1}^{X,\mathrm{PD}}$ with $(\alpha_A, \alpha_H) = (0.20, 0.10)$, The result is sent to 30-Hz position PD control.

Reward. Let $\kappa_\alpha(e) = \exp(-\alpha e)$. The reward is

$$r_t = r_t^H + r_t^O + r_t^{\mathrm{near}} + r_t^{\mathrm{multi}} + r_t^{\mathrm{lift}} + r_t^{\mathrm{eff}} - r_t^{\mathrm{reg}}. \tag{5}$$

The first two terms track the demonstration. Inspired by the contact shaping in [50], r_t^{near} and r_t^{multi} encourage fingertip approach and multi-finger contact. Together with action-smoothness and effort regularization, they improve sim-to-real transfer, while r_t^{lift} makes demonstrations containing grasp-and-lift motion easier to learn.

In particular, the imitation terms are

$$\begin{aligned}
r_t^H = g_t^L\big[&0.1\kappa_{40}(e_{w,p}) + 3\kappa_{1.91}(e_{w,R}) + \rho_t\textstyle\sum_{j\in\mathcal{G}} w_j\kappa_{\alpha_j}(e_j) \\
&+ 0.1\kappa_1(e_{w,v}) + 0.05\kappa_1(e_{w,\omega}) + 0.1\kappa_1(e_{J,v})\big],
\end{aligned} \tag{6}$$

$$\begin{aligned}
r_t^O = g_t^C\big[&5\kappa_{80}(e_{o,p}) + \kappa_3(e_{o,R}) + 0.1\kappa_1(e_{o,v}) \\
&+ 0.1\kappa_1(e_{o,\omega})\big].
\end{aligned} \tag{7}$$

A hat denotes a reference, and all translational errors use the FR3 base frame. For wrist or object $x \in \{w, o\}$, we use $e_{x,p} = \|\mathbf{p}_t^x - \hat{\mathbf{p}}_t^x\|_2$, $e_{x,v} = \|\mathbf{v}_t^x - \hat{\mathbf{v}}_t^x\|_1/3$, geodesic rotation error $e_{x,R}$, and $e_{x,\omega} = \|\boldsymbol{\omega}_t^x - \hat{\boldsymbol{\omega}}_t^x\|_1/3$. For MANO group $\mathcal{G}_j$, e_j is the mean pointwise L_2 position error and $e_{J,v}$ the mean absolute Cartesian velocity error over all MANO points. In the order thumb, index, middle, ring, pinky, level-1, and level-2, $\mathbf{w} = (0.9, 0.8, 0.75, 0.6, 0.6, 0.5, 0.3)$ and $\boldsymbol{\alpha} = (100, 90, 80, 60, 60, 50, 40)$. We set $\rho_t = 2$ in free space and 0.75 during contact.

Let d_i and F_i be elastomer fingertip–object distance and contact force, c_t^{ref} the number of reference fingertip contacts, and $c_t = \sum_i \mathbb{I}[F_i > 0.1\,\mathrm{N},\ d_i \leq 0.015\,\mathrm{m}]$ the rollout contact count. The binary gates are $g_t^C = \mathbb{I}[c_t^{\mathrm{ref}} > 0 \vee c_t > 0]$ and $g_t^L = 1 - \mathbb{I}[\hat{h}_t - h_0 > 0.05 \wedge h_t - h_0 \leq 0.05]$. Thus, g_t^C activates object tracking upon reference or rollout contact, whereas $g_t^L = 0$ suppresses dense hand tracking when a reference lift is not reproduced. Define $g_t^{\mathrm{ref},3} = \mathbb{I}[c_t^{\mathrm{ref}} \geq 3]$ and $g_t^{\mathrm{ref},+} = \mathbb{I}[c_t^{\mathrm{ref}} > 0]$. The two contact regularizers are $r_t^{\mathrm{near}} = 0.5 g_t^{\mathrm{ref},3}\sum_i \exp(-([d_i - 0.015]_+/0.020)^2)$ and $r_t^{\mathrm{multi}} = g_t^{\mathrm{ref},+}\{1.25 \text{ if } c_t = 4; 2.5 \text{ if } c_t = 5; 0 \text{ otherwise}\}$. We further use $r_t^{\mathrm{lift}} = 3\mathbb{I}[\hat{h}_t - h_0 > 0.05]\mathbb{I}[h_t - h_0 > 0.05]$ and $r_t^{\mathrm{eff}} = 0.5\kappa_{10}(P_t)$. Here h_t, $\hat{h}_t$, and h_0 are current, reference, and placed object heights, and $P_t = \sum_{k\in H}|\tau_{t,k}\dot{q}_{t,k}|$ is the hand-joint mechanical power. Finally, r_t^{reg} penalizes consecutive arm/hand raw-action differences with weights $(0.05, 0.025)$, joint velocities normalized by 15% of their limits with weights $(0.10, 0.04)$, and normalized torque above 20% of the effort limit with weight 0.2.

Policy optimization. We optimize with SAPG [12], which partitions parallel rollouts across Proximal Policy Optimization (PPO) agents with different exploration settings and aggregates their experience into a shared policy update. We pair this exploration mechanism with contact- and lift-aware rewards to guide learning from noisy multi-trajectory references.

Domain randomization. At every episode, we randomize arm and hand PD gains, object mass, object, table, and fingertip friction, and per-step actuation latency. We also corrupt the actor's object-pose observation while retaining clean simulator state for the privileged critic.

Episode initialization and termination. During training, we sample a reference frame and initialize the arm and object consistently with it. We retain the retargeted wrist pose but reset all 22 hand joints to an open configuration. Directly using the retargeted finger pose can initialize hand links inside the object or table, producing large depenetration impulses and unstable learning. An episode succeeds at the end of the reference and terminates early for object/hand tracking failure or exceeding joint speed thresholds.

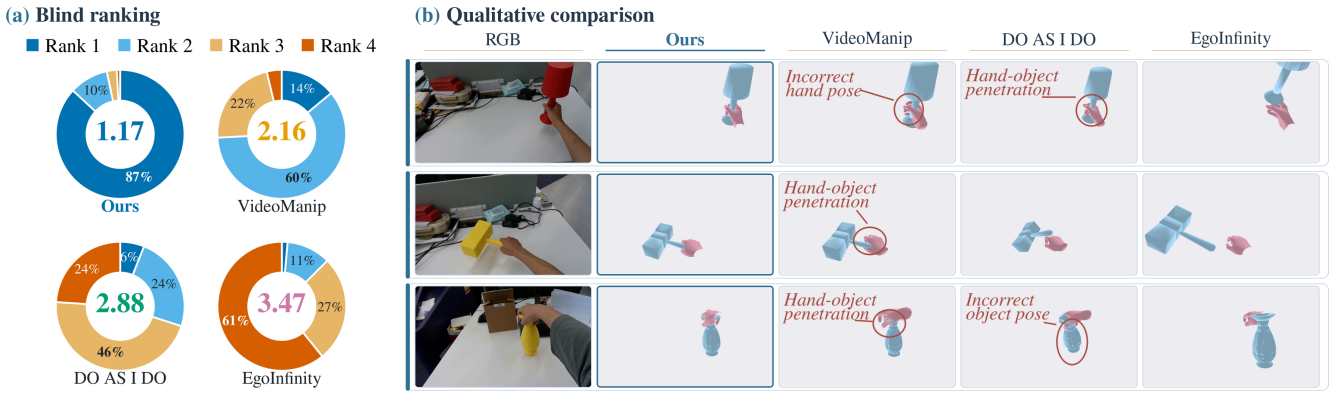

Fig. 4. Blind ranking and qualitative comparison on generated videos. (a) Rank distributions across 480 judgments per method; centers report mean rank (lower is better). (b) Representative frames using identical rendering, where our method better preserves object pose and hand–object placement.

TABLE I

HOI RECONSTRUCTION BENCHMARK AND COMPONENT ABLATIONS.

Method	**H2O (40)** · 150 frames/clip · 1280×720 px					**HO-Cap (40)** · 150 frames/clip · 640×480 px				
	ADD-S↓	**MRRPE↓**	**CDev↓**	**Avg. Rank↓**	**Time↓**	**ADD-S↓**	**MRRPE↓**	**CDev↓**	**Avg. Rank↓**	**Time↓**
Ours	**4.82**	**84.82**	**91.29**	**1.67**	12.33	**3.88**	**54.87**	**41.36**	**1.67**	6.73
w/o depth align.	5.14	86.97	**85.97**	**2.33**	–	6.01	57.04	**39.34**	**2.67**	–
w/o joint opt.	**5.10**	127.56	143.54	3.67	–	**4.02**	102.08	112.43	4.33	–
VideoManip	11.62	166.20	184.89	6.00	**5.97**	5.09	84.75	100.59	4.00	**4.72**
EgoInfinity	5.47	102.96	149.00	4.33	**11.12**	7.73	81.52	118.32	5.00	**5.79**
DO AS I DO	5.52	**71.83**	95.88	3.00	114.93	14.50	**46.53**	78.96	3.33	112.40

3) Policy Distillation: Direct training on the full reference set remains difficult under limited compute, whereas per-trajectory training [30] scales poorly. We therefore train at most two multi-skill experts per object category, each covering about 40 source trajectories (roughly 200 after augmentation). Each expert is obtained either by training from scratch on all trajectories of that category or, when this is less effective, by training on the full multi-object set and then fine-tuning on the target category. We distill these experts into a single policy through behavior cloning, then fine-tune the distilled policy with DAgger [51] to mitigate distribution shift.

IV. EXPERIMENTS

A. Ablation and Benchmark for HOI Reconstruction

Setup. We evaluate reconstruction in two settings. For generated videos without 3-D ground truth, four reviewers rank anonymized outputs of our method and three baselines on 120 clips sampled uniformly from 2,500 generations. Ties are allowed and method order is randomized, yielding 480 rankings per method. For quantitative evaluation, we use 40 fixed-view RGB-D clips from each of H2O [52] and HO-Cap [53], both with 3-D annotations. We compare the full method, its ablations, and the baselines using three metrics.

Let $\widehat{T}_t, T_t^*$ denote the predicted and ground-truth object-to-camera transforms and $\widehat{\mathbf{w}}_t, \mathbf{w}_t^*$ the corresponding wrist roots. ADD-S [54] measures closest-point object alignment as $\text{ADD-S}_t = |\mathcal{V}|^{-1} \sum_{\mathbf{v}\in\mathcal{V}} \min_{\mathbf{u}\in\mathcal{V}} \|\widehat{T}_t(\mathbf{v}) - T_t^*(\mathbf{u})\|_2$, where $\mathcal{V}$ contains at most 500 sampled CAD vertices. MRRPE_{ro} [55] measures error in the wrist-to-object-center vector, $\text{MRRPE}_{ro,t} = \|(\widehat{\mathbf{w}}_t - \widehat{T}_t(\bar{\mathbf{v}})) - (\mathbf{w}_t^* - T_t^*(\bar{\mathbf{v}}))\|_2$, where $\bar{\mathbf{v}}$ is the mesh centroid. CDev measures preservation of ground-truth contacts: we match each MANO vertex to its nearest CAD vertex, retain pairs within 3 mm, and average their distances under the predicted hand and object poses. We report ADD-S in centimeters and MRRPE_{ro}/CDev in millimeters. Avg. Rank averages the three metric ranks and excludes runtime on a single RTX 5880 (minutes/clip).

Baselines and ablations. We compare three baselines. *VideoManip* [56] reconstructs metric hand–object trajectories from monocular video and refines the hand pose using predicted contacts. *EgoInfinity* [57] calibrates hand and object estimates into a shared metric frame and refines object trajectories using inferred interaction states. *DO AS I DO* [5] combines world-space hand tracking with SAM3-based [58] object tracking and depth-based hand–object alignment. In ablation studies, *w/o depth align.* passes raw known-intrinsics MoGe-2 depth to FoundationPose and the joint optimizer, while *w/o joint opt.* removes Joint Hand–Object Optimization (Eq. (1)) and retains the independent FoundationPose object and WiLoR MANO estimates. All other components remain unchanged. Because *w/o depth align.* uses no measured depth, it also provides a comparison matched to the monocular baselines with respect to sensing modality; all methods are adapted to use the shared object mesh.

Results. Human raters place our method first in 417 of 480 judgments (87%), with a mean rank of 1.17 versus 2.16 for VideoManip, 2.88 for DO AS I DO, and 3.47 for EgoInfinity (Fig. 4(a)). The qualitative examples likewise preserve the observed object orientation and grasp placement, whereas the baselines show displacement, detachment, or hand–object penetration (Fig. 4(b)). Across H2O and HO-Cap, our method achieves the best Avg. Rank (1.67), the lowest ADD-S (4.82 cm on H2O and 3.88 cm on HO-

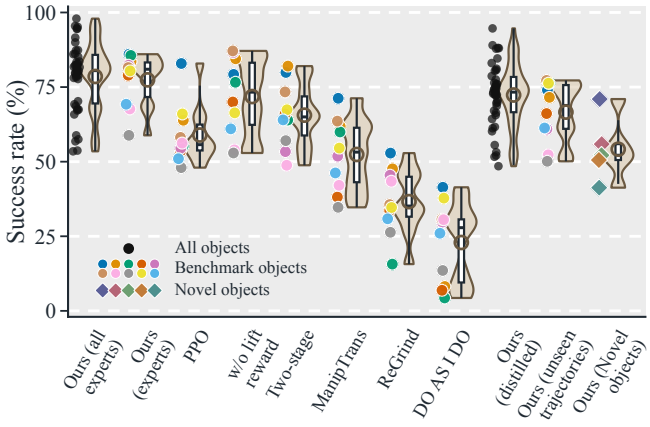

Fig. 5. Success rate statistics. Black circles, colored circles, and diamonds denote 42 training, ten benchmark, and five novel objects, respectively. Violins show distributions; boxes, bars, whiskers, and hollow circles mark interquartile ranges, medians, $1.5\times$IQR, and object-level means. Higher is better.

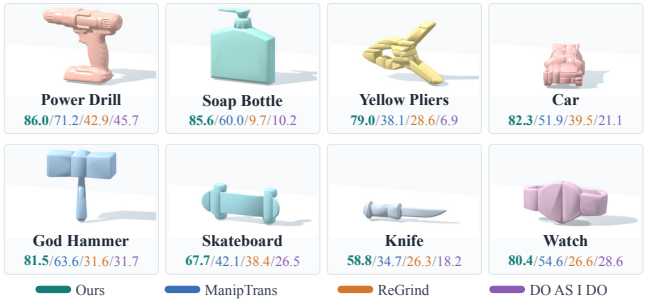

Fig. 6. Examples of manipulated objects. Values are the per-object success rates (%) of the methods in Fig. 5, in the same order. Meshes are normalized for visualization.

Cap), and second-best relative-geometry metrics (Table I). DO AS I DO uses the reconstructed hand as a metric anchor for per-frame object translation [5]. MRRPE$_{ro}$ naturally favors this hand-referenced alignment because it measures only wrist-to-object displacement and cancels absolute-depth errors shared by the hand and object. Its higher ADD-S and CDev, together with the human evaluation, show that this structural advantage under MRRPE$_{ro}$ does not imply better overall 3-D quality.

The ablations separate the roles of the two reconstruction stages. Even without depth alignment, our method has a better Avg. Rank than every baseline on both H2O (2.33) and HO-Cap (2.67). Its main degradation relative to the full method is in ADD-S, particularly on HO-Cap, showing that the overall quantitative gain does not arise solely from the measured depth anchor. Joint optimization markedly improves MRRPE$_{ro}$ and CDev on both datasets, confirming better relative hand–object geometry while preserving absolute object placement.

Remark: **Usable reference yield.** Of the 2,500 generated clips, 83% passed the reconstruction quality check, yielding about 2,000 usable references. As a point of reference, DO AS I DO's audit of 2,000 in-the-wild 100 Days of Hands clips found meaningful HOI in 187 (9.4%) and reconstructable HOI in 83 (4.2%) [5]. The criteria differ and the two sources are complementary, but the rates suggest that image-and-language conditioning can reduce human effort required to screen out irrelevant or unreconstructable videos.

B. Ablation and Benchmark for RL Formulation and Optimization

Setup. Fig. 5 reports the expert policies on all 42 training objects and a 10-object benchmark, together with the ablations, baselines, and distilled policy. The latter is evaluated on all training objects, 300 unseen trajectories of the benchmark objects, and five novel objects. Some benchmark objects are illustrated in Fig. 6. Benchmark methods share the evaluation reference set, with matched training budgets for learned policies. For each method–object

TABLE II

TRACKING ERRORS OVER INTENT-EXECUTING TRAJECTORIES.

Method	Pos. (mm) ↓	Rot. (°) ↓	Hand (mm) ↓
Ours	**10.2**	17.4	**36.2**
PPO	11.6	26.4	39.3
w/o lift reward	11.4	**13.3**	39.1
Two-stage	23.6	25.8	42.6
ManipTrans	25.8	23.6	39.7
ReGrind	17.7	14.0	76.8
DO AS I DO (floating)	35.3	7.1	36.9

experiment, we report success rate from 20,000 evaluation rollouts, each starting from the first reference frame. A trajectory is successful if its rollout reaches the final reference without exceeding a 4 cm object-position error, a 30° object-rotation error, or hand-keypoint thresholds of 6/6/8/10/12/12 cm for the thumb/index/middle/ring–pinky/level-1/level-2 groups. Tracking errors are averaged over *intent-executing* segments, i.e., after contact and a 5 cm lift until the first drop below 2.5 cm for grasp-and-move/pose-adjustment trajectories, and after first contact for push-and-pull trajectories, with segments shorter than eight steps discarded. Across all operation types, we report object-position error $\|\mathbf{p}_t - \hat{\mathbf{p}}_t\|_2$, geodesic rotation error $2\arccos(|\mathbf{q}_t^\top \hat{\mathbf{q}}_t|)$, and mean Euclidean error over 27 non-wrist hand keypoints.

Baselines and ablations. We compare with: 1) *PPO*, which replaces SAPG while keeping the remaining formulation of our method; 2) *w/o lift reward*, which removes r_t^{lift}; 3) *Two-stage*, our method with a hand imitation stage followed by a residual policy, which is similar to ManipTrans [27]; 4) *ManipTrans*, a two-stage method that learns hand-only trajectory imitation followed by a residual policy; 5) *ReGrind* [31], which uses interaction-aware retargeting to provide nominal targets for a residual PPO policy with object-keypoint and wrist tracking; and 6) *DO AS I DO* [5], a per-reference MPPI-style dynamics-aware retargeting method that directly optimizes an (open-loop) action sequence. The residual stages of ManipTrans and the ReGrind policy are originally optimized per reference but can be adapted to our multi-trajectory setting. We adapt both to our arm–hand embodiment and dynamics (including domain randomization) and align their observation spaces with ours. For DO AS I DO, we retain its floating-wrist formulation and align some of the dynamics parameters like hand PD gains and joint limits.

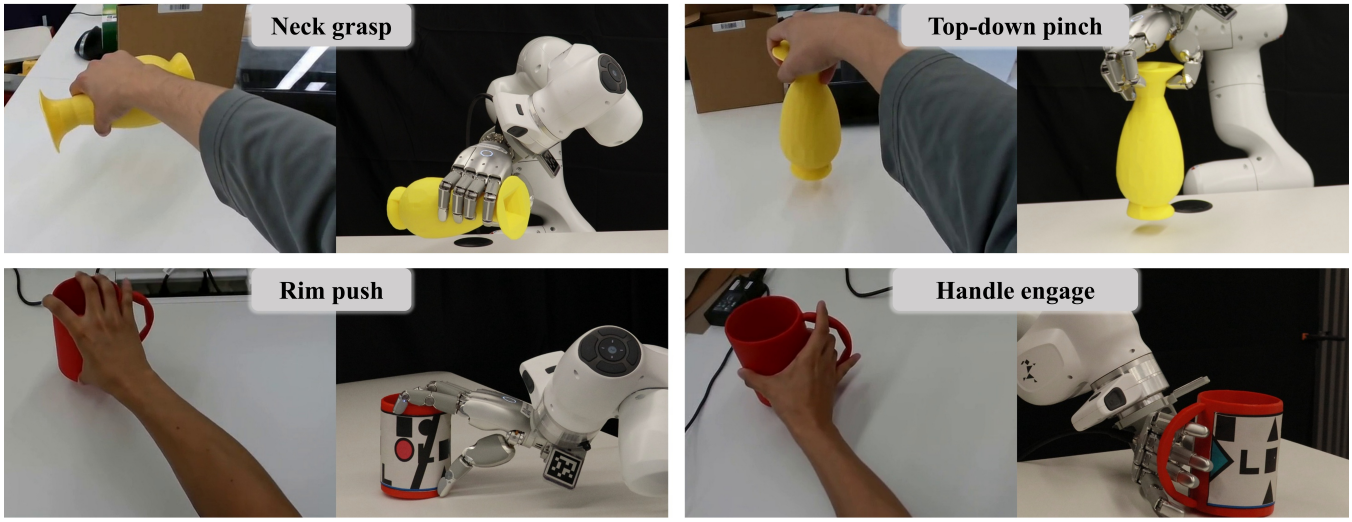

Fig. 7. Generated human references and corresponding real-world executions. All shown reference trajectories are unseen during policy training. The evaluated tasks comprise jar-neck pose adjustment, jar top-down grasp-and-move, mug-rim pushing, and mug-handle pulling.

Results. Across all 42 training objects, our expert policies reach a 78.6% macro mean and 80.5% median (Fig. 5). Distillation reduces the macro mean and median by only 4.2 and 6.0 percentage points, respectively. The distilled controller further reaches $66.6/68.8\%$ and $54.2/52.2\%$ mean/median on unseen reference trajectories and novel objects, respectively, showing transfer without retraining despite the substantial variation in manipulation difficulty. On the 10-object benchmark, Ours (experts) achieves the highest mean and median, 77.4% and 81.0%. Replacing SAPG with PPO, removing the lift reward, or switching to two-stage training reduces success on $10/10$, $7/10$, and $10/10$ objects, respectively. We sometimes observe faster early learning with two-stage training, but hypothesize that, when reconstructed hand–object relations are noisy, its first stage can impose suboptimal hand trajectories that restrict later-stage RL exploration. We note that ManipTrans was instead developed with cleaner motion-capture data [27], which explains this design choice. Our two-stage variant nevertheless exceeds ManipTrans on every benchmark object by 13.2 points on average, indicating that the reward formulation and optimizer together improve performance under the same training structure. Together with the lift-reward and two-stage ablations that retain SAPG, these results support the joint choice of reward formulation, training structure, and optimizer for noisy multi-trajectory tracking. ReGrind's lower success may arise because interaction-aware retargeting propagates reconstruction noise into distorted nominal targets, as well as its simpler observation and reward design provide less guidance for RL exploration.

Table II reports tracking errors computed over intent-executing segments. Our method achieves the lowest object-position and hand-keypoint errors among the arm–hand methods. We note that DO AS I DO reports a lower rotation error, while this number is not directly comparable since we find that DO AS I DO can not successfully plan a large number of pose-adjust trajectories that are excluded as not meeting the intent-executing criteria.

TABLE III

REAL-WORLD SUCCESS ON UNSEEN TRAJECTORIES (10 TRIALS/TASK).

Task	Jar neck	Jar top	Mug rim	Mug handle	Overall
Type	Pose-adjust	Grasp-and-move	Push-and-pull	Push-and-pull	–
Success	6/10	6/10	7/10	8/10	**27/40**

C. Real-World Experiments

Real-World Setup and Onboard Perception. An Intel RealSense D455 observes the workspace from a viewpoint largely free of arm occlusion and provides online object poses through FoundationPose [45]. Another Intel RealSense D435 is used only to capture the scene image that conditions video generation, which is removed before execution to avoid collision with the robot and does not participate in closed-loop perception. Because deployment is sensitive to translational calibration error, beyond typical hand-eye calibration, we command the robot rigidly grasp an object and use the known end-effector pose to estimate a shared constant bias in the FoundationPose translation across objects, analogous to the Tsai–Lenz calibration method.

Results. Here we present the results of evaluating the distilled controller on 40 unseen video plans. The four tasks are jar-neck pose adjustment, jar top-down grasp-and-move, mug-rim pushing, and mug-handle pulling, with each task having 10 trials. Each plan is conditioned on a scene image and a new language instruction. Success follows the same trajectory-completion criteria and tracking-error thresholds as in Sec. IV-B, using object poses from FoundationPose and hand keypoints computed by forward kinematics (FK) from measured robot joint positions. The controller achieves 27/40 successes overall, with per-task results reported in Table III. Representative failures are discussed in Sec. V. As shown in Fig. 7, the jar-neck task uses a small-diameter power grasp that leaves the opening accessible, whereas the top-down task uses a multi-finger pinch at the opening. These

grasp choices are part of the interaction specified by the video plan, highlighting the importance of retaining both hand and object motion in the reference compared to object-only ones [19].

V. DISCUSSION AND CONCLUSION

1) Failure Modes and Potential Improvements: **Training-time failures.** Some generated interactions appear plausible but are difficult for a robotic hand. For example, humans can lift large objects through palm friction far from force closure, whereas such motions are hard to learn in simulation. Our method and the baselines also struggle with flat or thin objects when the grasp lies near the table. Task-specific RL policies [50] may provide useful action supervision or motion priors [7] to address these issues.

Deployment failures. Contact transitions are sensitive to the sim-to-real gap and can cause large tracking deviations. Although the controller shows basic recovery and retry behavior, it often fails after leaving the reference. Embedding training in a more generalized goal-conditioned MDP beyond merely tracking could provide more diverse recovery data.

2) Towards Versatile Dexterous Controllers: Our skill set does not systematically cover in-hand manipulation, which appears only incidentally in generated videos. Video references are less reliable for such contact-rich motion because reconstruction noise obscures subtle finger–object motion. We therefore aim to combine multiple data sources in a unified controller for everyday manipulation of single rigid objects. Nevertheless, the grasping and non-prehensile skills learned here demonstrate the potential of generated video and mark a concrete step toward versatile dexterous control.

ACKNOWLEDGMENT

We thank Kaihan Chen for assistance with implementing the baseline methods, and Ruoqu Chen and Kechun Xu for early suggestions on key technical design choices. We also thank Junxiao Lin, Yipeng Pan, Xingyu Ji, and Mingjie Zhou for conducting the human preference evaluation of our reconstruction pipeline.

REFERENCES

[1] A. N. Meltzoff, "Imitation of televised models by infants," *Child Dev.*, vol. 59, no. 5, pp. 1221–1229, 1988.

[2] A. Whiten, "The burgeoning reach of animal culture," *Science*, vol. 372, no. 6537, p. eabe6514, 2021.

[3] C. Wang, L. Fan, J. Sun, *et al.*, "MimicPlay: Long-horizon imitation learning by watching human play," in *Proc. 7th Conf. Robot Learn. (CoRL)*, vol. 229, pp. 201–221, 2023.

[4] S. Kareer, D. Patel, R. Punamiya, *et al.*, "EgoMimic: Scaling imitation learning via egocentric video," *arXiv:2410.24221*, 2024.

[5] B. Paliwal, H. Etukuru, W. Liang, P. Abbeel, N. M. M. Shafiullah, and J. Malik, "Do as I do: Dexterous manipulation data from everyday human videos," *arXiv:2606.19333*, 2026.

[6] R. Punamiya, S. Kareer, Z. Liu, *et al.*, "EgoVerse: An egocentric human dataset for robot learning from around the world," in *Robot.: Sci. Syst. (RSS)*, 2026.

[7] Z. Luo, Y. Yuan, T. Wang, *et al.*, "SONIC: Supersizing motion tracking for natural humanoid whole-body control," *Sci. Robot.*, vol. 11, no. 117, p. eaed4592, 2026.

[8] Team Seedance, D. Chen, L. Chen, *et al.*, "Seedance 2.0: Advancing video generation for world complexity," *arXiv:2604.14148*, 2026.

[9] V. Makoviychuk, L. Wawrzyniak, Y. Guo, *et al.*, "Isaac Gym: High-performance GPU-based physics simulation for robot learning," *arXiv:2108.10470*, 2021.

[10] Z. Chen, S. Chen, E. Arlaud, I. Laptev, and C. Schmid, "ViViDex: Learning vision-based dexterous manipulation from human videos," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2025.

[11] T. G. W. Lum, O. Y. Lee, C. K. Liu, and J. Bohg, "Crossing the human–robot embodiment gap with sim-to-real RL using one human demonstration," in *Proc. Conf. Robot Learn. (CoRL)*, 2025.

[12] J. Singla, A. Agarwal, and D. Pathak, "SAPG: Split and aggregate policy gradients," in *Proc. 41st Int. Conf. Mach. Learn. (ICML)*, vol. 235, pp. 45759–45772, 2024.

[13] Y. Du, M. Yang, B. Dai, *et al.*, "Learning universal policies via text-guided video generation," in *Adv. Neural Inf. Process. Syst.*, 2023.

[14] H. Bharadhwaj, D. Dwibedi, A. Gupta, *et al.*, "Gen2Act: Human video generation in novel scenarios enables generalizable robot manipulation," in *Proc. Conf. Robot Learn. (CoRL)*, 2025.

[15] S. Ye, Y. Ge, K. Zheng, *et al.*, "World action models are zero-shot policies," *arXiv:2602.15922*, 2026.

[16] L. Li, Q. Zhang, Y. Luo, *et al.*, "Causal world modeling for robot control," *arXiv:2601.21998*, 2026.

[17] J. Liang, R. Liu, E. Ozguroglu, *et al.*, "Dreamitate: Real-world visuomotor policy learning via video generation," *arXiv:2406.16862*, 2024.

[18] B. Chen, T. Zhang, H. Geng, *et al.*, "Large video planner enables generalizable robot control," *arXiv:2512.15840*, 2025.

[19] Y. Kuang, S. Park, K. Fragkiadaki, and S. Tulsiani, "Dex4D: Task-agnostic point track policy for sim-to-real dexterous manipulation," *arXiv:2602.15828*, 2026.

[20] J. Ni, Z. Wang, W. Lin, *et al.*, "From generated human videos to physically plausible robot trajectories," *arXiv:2512.05094*, 2025.

[21] J. Chen, Z. Wang, F. Jia, *et al.*, "Imagine2Real: Towards zero-shot humanoid–object interaction via video generative priors," *arXiv:2605.22272*, 2026.

[22] T. Xie, H. Zhang, J. Park, *et al.*, "GRAIL: Generating humanoid loco-manipulation from 3D assets and video priors," *arXiv:2606.05160*, 2026.

[23] D. Kahneman, P. Egan, *et al.*, *Thinking, fast and slow*, vol. 1. Farrar, straus and Giroux New York, 2011.

[24] H. Gupta, G. Shi, and W. Yuan, "LUCID: Learning embodiment-agnostic intent models from unstructured human videos for scalable dexterous robot skill acquisition," *arXiv:2606.11628*, 2026.

[25] W. Xing, Z. Zhang, J. Yin, Y. Ye, R. Li, Y. Yao, S. Li, X. Zhu, and K. Zhang, "Decoupled Dexterity: Learning Intent from Demonstrations and Reflexes from Interaction," tech. rep., Sharpa, 2026.

[26] S. Zhao, X. Zhu, Y. Chen, *et al.*, "DexH2R: Task-oriented dexterous manipulation from human to robots," *arXiv:2411.04428*, 2024.

[27] K. Li, P. Li, T. Liu, Y. Li, and S. Huang, "ManipTrans: Efficient dexterous bimanual manipulation transfer via residual learning," in *Proc. CVPR*, 2025.

[28] M. Zhao, Y. Hou, D. Fox, Y. Narang, A. Mandlekar, and S. Song, "DexMachina: Functional retargeting for bimanual dexterous manipulation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2026.

[29] X. Zhu, Z. Liu, S. Jain, *et al.*, "Learning dexterous manipulation using contact wrench guidance from human demonstration," *arXiv:2607.00033*, 2026.

[30] X. Liu, J. Adalibieke, Q. Han, Y. Qin, and L. Yi, "DexTrack: Towards generalizable neural tracking control for dexterous manipulation from human references," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2025.

[31] Y. Feng, N. Leung, J. Wang, *et al.*, "A minimalist retargeting-guided reinforcement learning recipe for dexterous manipulation," *arXiv:2607.11874*, 2026.

[32] P. Li, Z. Chen, Y. Wu, P. Wei, Y. Li, T. Wang, J. Shi, M. Yu, B. Jia, S.-c. Zhu, *et al.*, "Towards human-level dexterous teleoperation," *arXiv preprint arXiv:2607.11481*, 2026.

[33] Y. Qin, Y.-H. Wu, S. Liu, *et al.*, "DexMV: Imitation learning for dexterous manipulation from human videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 570–587, 2022.

[34] J. Hsieh, K.-H. Tu, K.-H. Hung, and T.-W. Ke, "DexMan: Learning bimanual dexterous manipulation from human and generated videos," *arXiv:2510.08475*, 2025.

[35] K. Chen, Y. Shao, H. Ji, X. Yang, and Y. Mu, "V2P-Manip: Learning dexterous manipulation from monocular human videos," *arXiv:2606.16436*, 2026.

[36] R. Chen, F. Ruan, L. Cao, Z. Wang, B. Xu, S. Tong, J. Liu, M. Pei, C. Zhang, W. Xing, *et al.*, "Dex-x: Learning visual-tactile dexterous manipulation from human videos with simulated interaction," *challenge*, vol. 1, no. 29, p. 30.

[37] Y. Wang, R. Yu, H. W. Tsui, *et al.*, "Learning generalizable hand–object tracking from synthetic demonstrations," *arXiv:2512.19583*, 2025.

[38] J. Adalibieke, Q. Han, X. Liu, Y. Qin, and L. Yi, "AdaDexTrack: Dynamic modulation for adaptive and generalizable dexterous manipulation tracking," in *Proc. CVPR*, 2026.

[39] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 1–17, 2017.

[40] S. Hampali, M. Rad, M. Oberweger, and V. Lepetit, "HOnnotate: A method for 3D annotation of hand and object poses," in *Proc. CVPR*, pp. 3196–3206, 2020.

[41] Z. Fan, M. Parelli, M. E. Kadoglou, *et al.*, "HOLD: Category-agnostic 3D reconstruction of interacting hands and objects from video," in *Proc. CVPR*, pp. 494–504, 2024.

[42] R. Wang, S. Xu, Y. Dong, *et al.*, "MoGe-2: Accurate monocular geometry with metric scale and sharp details," *arXiv:2507.02546*, 2025.

[43] N. Ravi, V. Gabeur, Y.-T. Hu, *et al.*, "SAM 2: Segment anything in images and videos," *arXiv:2408.00714*, 2024.

[44] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[45] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "FoundationPose: Unified 6D pose estimation and tracking of novel objects," in *Proc. CVPR*, pp. 17868–17879, 2024.

[46] R. A. Potamias, J. Zhang, J. Deng, and S. Zafeiriou, "WiLoR: End-to-end 3D hand localization and reconstruction in-the-wild," in *Proc. CVPR*, pp. 12242–12254, 2025.

[47] A. Darkhalil, D. Damen, and D. Fouhey, "Improving and evaluating hand–object interaction detection," *arXiv:2606.17384*, 2026.

[48] Sharpa Robotics, "SharpaWave: A dexterous robotic hand," 2025. [Online]. Available: https://www.sharpa.com/pages/wave.

[49] Franka Robotics, "Franka Research 3," 2022. [Online]. Available: https://franka.de/franka-research-3.

[50] H. Zhang, Z. Wu, L. Huang, S. Christen, and J. Song, "Robust dexterous grasping of general objects," in *Proc. 9th Conf. Robot Learn.*, vol. 305, pp. 3035–3050, 2025.

[51] S. Ross, G. Gordon, and J. A. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proc. 14th Int. Conf. Artif. Intell. Stat. (AISTATS)*, vol. 15, pp. 627–635, 2011.

[52] T. Kwon, B. Tekin, J. Stuehmer, *et al.*, "H2O: Two hands manipulating objects for first person interaction recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021.

[53] J. Wang, Q. Zhang, Y.-W. Chao, *et al.*, "HO-Cap: A capture system and dataset for 3D reconstruction and pose tracking of hand-object interaction," in *Adv. Neural Inf. Process. Syst.: Datasets and Benchmarks*, 2025.

[54] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes," in *Robot.: Sci. Syst. (RSS)*, 2018.

[55] Z. Fan, O. Taheri, D. Tzionas, *et al.*, "ARCTIC: A dataset for dexterous bimanual hand-object manipulation," in *Proc. CVPR*, 2023.

[56] H. Chen, T. Dong, T. Wu, *et al.*, "Dexterous manipulation policies from RGB human videos via 3D hand-object trajectory reconstruction," *arXiv:2602.09013*, 2026.

[57] G. Wang, K. Ren, A. Morgan, *et al.*, "EgoInfinity: A web-scale 4D hand-object interaction data engine for any-view robot retargeting and video-to-action robot learning," *arXiv:2606.17385*, 2026.

[58] SAM 3D Team, X. Chen, F.-J. Chu, *et al.*, "SAM 3D: 3Dfy anything in images," in *Proc. CVPR*, 2026.