DeVI: Physics-based Dexterous Human-Object Interaction via Synthetic Video Imitation

Anonymous authors
Paper under double-blind review

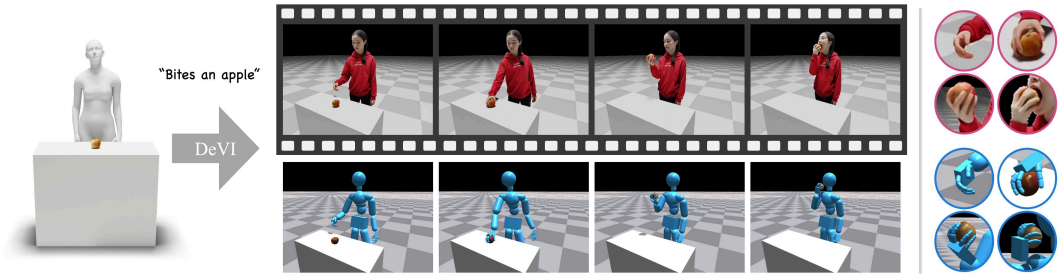


Figure 1: **DeVI.** Given a physics environment with 3D human and objects along with an interaction text prompt, our method, DeVI, generates a physically plausible dexterous human-object interaction motion by using a video diffusion model as an interaction-aware motion planner.

Abstract

Physics-based character control for dexterous Human-Object Interaction (HOI) relies heavily on motion-capture data, which limits its applicability beyond captured scenarios. Meanwhile, video generative models can synthesize realistic HOI videos for unseen objects and scenarios directly from text prompts, but 3D interactions learned from such synthetic data remain kinematic without ensuring physical plausibility. Imitating these videos with physics-based control is a promising direction, yet unreliable object tracking and hand-object misalignment make it difficult to obtain accurate imitation targets. We present DeVI (Dexterous Video Imitation), a novel framework that learns physics-based character control for dexterous HOI by imitating text-conditioned synthetic videos. To address unreliable 3D object motion reconstruction, we introduce hybrid imitation targets that combine reconstructed 3D human motion with tracked 2D object trajectories, bypassing the explicit reconstruction of the 3D object motion. Our Visual HOI Alignment further refines the human reference to align with both the generated video and the initial 3D object configuration, making it suitable for physical interaction. Extensive experiments show that DeVI's hybrid representation is more effective than those used by the baselines and that Visual HOI Alignment improves reconstruction quality and imitation success. We further demonstrate physically plausible functional manipulation from text instructions, including target-aware interactions and articulated object manipulation, without requiring motion-capture demonstrations.

1 Introduction

Modeling physics-based human motion is essential for understanding realistic human behavior involving physical interaction with the environment, including dexterous object manipulation. Existing studies on human motion simulation (Peng et al., 2018; Luo et al., 2023; 2024a; Zhang et al., 2025b; Tessler et al., 2024; Peng et al., 2021; 2022; Tevet et al., 2025; Yuan et al., 2023) primarily focus on human motion alone, while the approaches addressing HOI in physical environments often overlook dexterous hand-object interactions involving finger articulation (Cui et al., 2024; 2025) or focus on specific domains such as sports (Luo et al., 2024b; Liu & Hodgins, 2018; Haarnoja et al., 2024; Hong et al., 2019). Recent approaches show promising results in imitating dexterous HOI (Wang et al., 2023; Wu et al., 2025; Lin et al., 2025) and in learning unified control policies (Wang et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2025a; Xu et al., 2025), but still rely on 3D HOI motion-capture (MoCap) demonstrations (Taheri et al., 2020; Li et al., 2023), whose limited coverage constrains the diversity of interactions they can learn. Since dexterous manipulation is difficult to capture at scale, extending physics-based character control to diverse objects and novel functional interactions remains challenging.

Meanwhile, video generative models (Yang et al., 2024b; Wan et al., 2025) enable zero-shot synthesis of high-fidelity HOI videos for unseen objects and scenarios from text prompts, offering rich knowledge of grasping and functional object use (*e.g.*, drinking coke, opening a microwave). Recent methods leverage such synthetic videos for HOI generation (Kim et al., 2025; Li & Dai, 2026), but the resulting interactions remain kinematic rather than being driven by physics-based control. Imitating these videos with existing physics-based HOI methods (Xu et al., 2025; Wang et al., 2023; 2025a) is a promising direction, but it requires both human and object motion in 3D. While human motion can be recovered via HMR (Human Mesh Recovery) (Shen et al., 2024; Wang et al., 2024b), as in human-only video imitation (He et al., 2025; Zhang et al., 2025a), 3D object motion reconstruction from synthetic videos (Patel et al., 2026; Kuang et al., 2026) remains unreliable due to monocular depth ambiguity. The resulting trajectory errors cause hand-object misalignment and hinder physics-based imitation (Fig. 4). Even with accurate human and object references, universal policies trained on limited 4D HOI datasets fail to generalize to such novel interactions (Fig. 10).

To address these challenges, we present **DeVI (Dexterous Video Imitation)**, a framework that learns physics-based character control for dexterous HOI by imitating text-conditioned synthetic videos, without requiring 3D HOI MoCap demonstrations. A key advantage of DeVI is that it learns physics-based control for novel functional interactions, from grasp acquisition to text-driven object use. To bypass unreliable 3D object motion reconstruction, we introduce *hybrid imitation targets* that combine reconstructed 3D human motion with 2D object tracks from a generated HOI video. Notably, we observe that precise alignment between the lifted 3D human motion and the initial 3D scene is crucial for physical interaction (Fig. 6). We therefore introduce *Visual HOI Alignment* to refine the human motion to match both the reference video and the initial 3D object configuration. The resulting hybrid targets are used to train a humanoid control policy via reinforcement learning (RL) with a hybrid tracking reward that combines 3D human and 2D object tracking.

We evaluate DeVI on both generated HOI videos with Internet-sourced objects (SketchFab) and 3D MoCap demonstrations from GRAB (Taheri et al., 2020). For video-based imitation, we compare our hybrid imitation targets against reconstructed 3D object motion (Patel et al., 2026; Kuang et al., 2026; Kim et al., 2025) and 4D HOI references (Xie et al., 2026b; Li & Dai, 2026). For MoCap-based imitation, we compare against methods that directly imitate 3D HOI motions (Wang et al., 2023; 2025a; Xu et al., 2025), demonstrating the benefits of our hybrid tracking formulation even when accurate 3D references are available. We further demonstrate the advantages of synthetic video imitation through diverse functional manipulation scenarios, including target-aware interactions and articulated object manipulation. Finally, we validate Visual HOI Alignment through ablation studies and comparisons with existing HMR methods (Wang et al., 2024b; 2025b; Shen et al., 2024).

In summary, our main contributions are as follows. (1) We introduce DeVI, a framework that learns physics-based character control for dexterous HOI by imitating text-conditioned synthetic videos. (2) We present hybrid imitation targets that combine reconstructed 3D human motion with 2D object tracks, bypassing explicit 3D object motion reconstruction. (3) We propose Visual HOI Alignment to refine reconstructed human motion to align with both the reference video and the initial 3D object. (4) We demonstrate the advantages of synthetic video imitation through various text-driven functional manipulation scenarios, including target-aware interactions and articulated object manipulation. Our code and results will be publicly released to support reproducibility.

2 Related Work

Video Generation for Robotic Manipulation. Recent advances in video generation motivate the use of generated videos as visual plans for robotic manipulation. Building on this, recent approaches translate visual predictions into robot control either by learning robot actions from video representations (Du et al., 2023; Wu et al., 2024; Wen et al., 2024; Feng et al., 2025) or by recovering explicit geometric motion, including 6D tool trajectories (Liang et al., 2024) and 6D object trajectories (Patel et al., 2026). Recent methods extend the paradigm to dexterous manipulation with robotic hands (Hu et al., 2025; Chen et al., 2025a; Kuang et al., 2026). These methods primarily

target robot task execution, while physics-based full-body HOI imitation must also include human motion, making it particularly sensitive to spatio-temporal alignment between 3D human motion and object trajectories. We address the issue through our hybrid imitation targets that combine 3D human motion with 2D object tracks, avoiding explicit 3D object trajectory reconstruction from the video.

Monocular HOI Reconstruction. Recovering 3D structure from 2D observations is a long-standing challenge in computer vision due to inherent depth ambiguity. Recent advances substantially improve monocular 3D reconstruction of scenes (Cao & de Charette, 2022; Yang et al., 2024a; Wang et al., 2024a), objects (Xu et al., 2024; Xiang et al., 2025; Team et al., 2025), and humans (Kanazawa et al., 2018; Shen et al., 2024; Wang et al., 2024b). However, jointly reconstructing a human interacting with an object is considerably more difficult, as it requires precise spatio-temporal alignment between the two. Early approaches rely on pre-defined object templates and hand-crafted contact heuristics (Zhang et al., 2020). Subsequent studies replace such heuristics with learning-based contact estimators (Xie et al., 2022) and extend the setting to a category-agnostic one (Dwivedi et al., 2025), yet remain restricted to single-frame reconstruction. Recent works extend HOI reconstruction to video-based tracking (Xie et al., 2023; 2025; 2026b), but remain limited to body-level interactions without modeling dexterous hand motion. To tackle the challenging nature of 4D HOI reconstruction with dexterous hands, we introduce Visual HOI Alignment, which aligns the reconstructed human motion with the generated video and the object in the initial 3D scene.

Physics-based HOI Motion Generation. Traditional studies (Peng et al., 2018) learn a humanoid control policy via RL to imitate a single MoCap reference motion. Since learning a separate policy per motion is expensive, several studies learn an integrated policy through adversarial motion priors (Peng et al., 2021; 2022; Tessler et al., 2023) that capture the distribution of motions in the dataset. To further generalize humanoid control across diverse motions, several approaches (Luo et al., 2023; 2024a; Tessler et al., 2024) train a unified policy imitating the given demonstrations. This line of work is extended to HOI, achieving single HOI imitation (Wang et al., 2023), interaction skill learning (Hassan et al., 2023; Xiao et al., 2024; Pan et al., 2024; Gao et al., 2024; Wang et al., 2025a; Zhao et al., 2025), and general HOI imitation (Xu et al., 2025). However, these approaches rely on 3D human-object motion references that are difficult to obtain at scale, while controllers trained on such MoCap data show limited generalization beyond their training distribution. A recent study (Lou et al., 2025) generates physically plausible HOI motion via a text-to-motion planner, but also shows limited generalization and starts from a pre-grasped state, skipping the important and challenging grasping stage. We mitigate this problem by using synthetic video as a motion planner that covers the full manipulation sequence, including approach, grasp, and release.

3 DEVI: DEXTEROUS VIDEO IMITATION

Given an initial scene consisting of a 3D human, represented by an SMPL-X (Pavlakos et al., 2019) pose, and a 3D object, our goal is to learn a policy π that imitates a text-conditioned synthetic video including dexterous HOI in a physics environment. Our core idea for imitating the synthetic video is to extract hybrid imitation targets from it, which include 3D human motion and 2D object trajectories, and to use them to train the physics-based character control policy.

In this paper, we focus on dexterous hand-object manipulation in tabletop scenarios. We initialize a scene $\mathcal{S} = \{\mathcal{H}, \mathcal{O}\}$ as detailed in Appendix A.1, where $\mathcal{H}$ denotes the human parameterized by SMPL-X, and $\mathcal{O}$ represents the object defined as follows:

$$\mathcal{H} = \{\boldsymbol{\beta}, \boldsymbol{\theta}^b, \boldsymbol{\phi}^b, \boldsymbol{\tau}^b, \boldsymbol{\theta}^h\}, \quad \mathcal{O} = \{\boldsymbol{\phi}^o, \boldsymbol{\tau}^o\}, \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^{10}$, $\boldsymbol{\theta}^b \in \mathbb{R}^{21\times3}$, $\boldsymbol{\theta}^h \in \mathbb{R}^{30\times3}$, $\boldsymbol{\phi}^b \in \mathrm{SO}(3)$, and $\boldsymbol{\tau}^b \in \mathbb{R}^3$ are the SMPL-X parameters of human shape, body pose, hand pose, global body root orientation, and global body root translation, respectively. The $\boldsymbol{\phi}^o \in \mathrm{SO}(3)$ and $\boldsymbol{\tau}^o \in \mathbb{R}^3$ are the object's global orientation and translation. We define $\mathcal{M}_O(\boldsymbol{\phi}^o, \boldsymbol{\tau}^o)$ as a mapping from the object state to the transformed object mesh. We use a similar function $\mathcal{M}_{\text{SMPLX}}(\mathcal{H})$ to obtain the SMPL-X mesh from the parameters. In this section, we consider the single-object case for simplicity, although our framework readily extends to multi-object.

We first synthesize an HOI video from the input scene following the prompt (Sec. 3.1). From the video, we obtain hybrid imitation targets (Sec. 3.2). Then, we present our hybrid tracking reward (Sec. 3.3) for policy learning. An overview of our method is shown in Fig. 2.

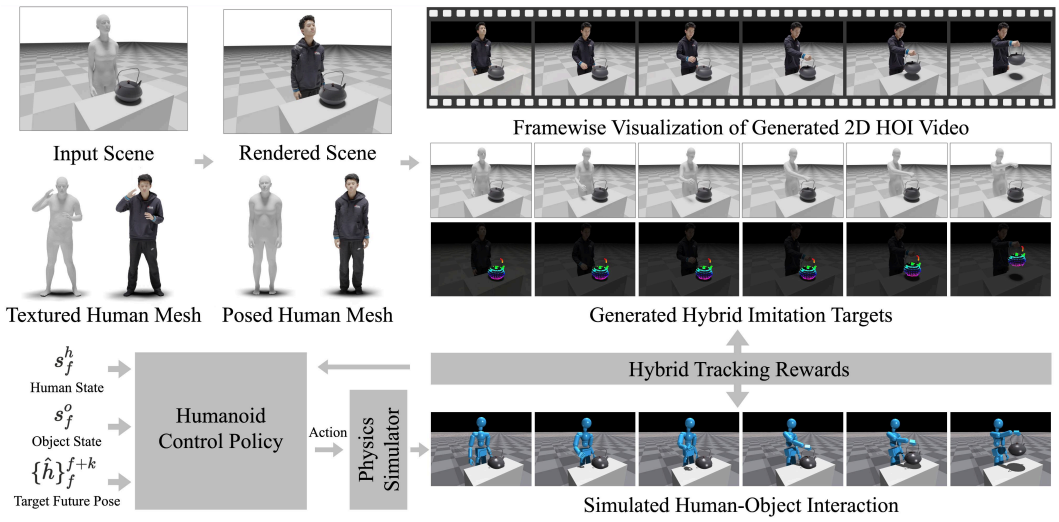


Figure 2: **Overview.** Given a scene with an SMPL-X (Pavlakos et al., 2019) human and object, we replace it with a deformed textured mesh and render to generate HOI video. Then hybrid imitation targets extracted from the video are used to train our humanoid control policy.

3.1 2D HOI Video Generation

As 2D video synthesis and 3D human mesh recovery perform better with realistic appearances, we replace the SMPL-X surface mesh with a textured human mesh $\mathcal{M}_{\text{Human}}(\mathcal{H})$ that matches the scale, pose, and location of $\mathcal{M}_{\text{SMPLX}}(\mathcal{H})$. In practice, we select textured human meshes from the THuman2.0 dataset (Yu et al., 2021) and deform them by automatic rigging with approximated joint offsets and skinning weights for linear blend skinning (LBS). Appendix A.2 details the rigging process. We then specify a camera with parameters Π and render the scene into 2D image space $\mathcal{I} = \Pi(\{\mathcal{M}_{\text{Human}}(\mathcal{H}), \mathcal{M}_O(\mathcal{O})\})$, together with a fixed table and background. Using a pre-trained image-to-video model (Wan et al., 2025) and the text prompt, we generate a 2D HOI video $\mathcal{V} = \{\mathcal{I}_t\}_{t=1}^F$ initialized from the rendered image $\mathcal{I}_1 = \mathcal{I}$.

3.2 Extracting Hybrid Imitation Targets

Directly using the generated video $\mathcal{V}$ for RL policy learning is challenging. Instead, we first extract hybrid imitation targets $\hat{\boldsymbol{g}}^{\text{hybrid}} = \{\hat{\boldsymbol{h}}, \hat{\boldsymbol{o}}\}$, which consist of estimated 3D human motion $\hat{\boldsymbol{h}}$ and 2D object trajectories $\hat{\boldsymbol{o}}$ from the video as follows:

$$\hat{\boldsymbol{h}} = \{\hat{\boldsymbol{J}}_t^b, \hat{\boldsymbol{\theta}}_t^b, \hat{\boldsymbol{J}}_t^h, \hat{\boldsymbol{\theta}}_t^h\}_{t=1}^F, \quad \hat{\boldsymbol{o}} = \{\hat{\boldsymbol{x}}_t\}_{t=1}^F. \tag{2}$$

Here, $\hat{\boldsymbol{J}}_t^b \in \mathbb{R}^{19\times 3}$ and $\hat{\boldsymbol{J}}_t^h \in \mathbb{R}^{32\times 3}$ denote the estimated 3D body and hand joint locations, and $\hat{\boldsymbol{\theta}}_t^b \in \mathbb{R}^{19\times 3}$ and $\hat{\boldsymbol{\theta}}_t^h \in \mathbb{R}^{32\times 3}$ represent the corresponding body and hand pose parameters in SMPL-X representation. The object target $\hat{\boldsymbol{o}}$ consists of tracked 2D points $\hat{\boldsymbol{x}}_t \in \mathbb{R}^{M\times 2}$ in image coordinates at time t, initialized from the first frame. These points correspond to M visible object vertices projected into the image using the camera Π. Our motivation for using 3D human pose with 2D object trajectories is that accurately recovering full 3D object pose from video remains challenging as shown in Fig. 4, and robust 3D spatial alignment between humans and objects is still an open problem. By providing these hybrid cues as reference targets, we allow the RL framework to infer a physically plausible solution that jointly imitates both signals. Below, we describe the extraction process for each component of the hybrid imitation target.

Initializing 3D Human Reference. For the human component, we apply an off-the-shelf monocular world-grounded human motion estimator (Shen et al., 2024), denoted as $\mathcal{F}^b$, along with a hand pose estimator (Pavlakos et al., 2024), denoted as $\mathcal{F}^h$, to the synthesized 2D video $\mathcal{V}$. These models provide estimates of 3D body motion and 3D hand poses, which we combine to reconstruct a unified SMPL-X human motion sequence with hand poses. Specifically,

$$\mathcal{F}^b(\mathcal{I}_t) = \{\boldsymbol{\beta}_t, \boldsymbol{\theta}_t^b, \boldsymbol{\phi}_t^b, \boldsymbol{\tau}_t^b\}, \quad \mathcal{F}^h(\mathcal{I}_t) = \{\boldsymbol{\theta}_t^h, \boldsymbol{\phi}_t^{lh}, \boldsymbol{\phi}_t^{rh}, \boldsymbol{\tau}_t^{lh}, \boldsymbol{\tau}_t^{rh}\}, \tag{3}$$

where $\boldsymbol{\beta}_t \in \mathbb{R}^{10}$, $\boldsymbol{\theta}_t^b \in \mathbb{R}^{21\times3}$, $\boldsymbol{\phi}_t^b \in \mathrm{SO}(3)$, and $\boldsymbol{\tau}_t^b \in \mathbb{R}^3$ are the estimated SMPL-X parameters for human shape, body pose, the root joint orientations in the world coordinate, and translation, respectively. The $\boldsymbol{\theta}_t^h \in \mathbb{R}^{30\times3}$ is a hand pose, $\boldsymbol{\phi}_t^{lh}, \boldsymbol{\phi}_t^{rh} \in \mathrm{SO}(3)$ are the root (wrist) orientations of each hand, and $\boldsymbol{\tau}_t^{lh}, \boldsymbol{\tau}_t^{rh} \in \mathbb{R}^3$ are the global translations of each hand. As the hand estimator typically provides more accurate hand locations and orientations, we refine the body pose by adjusting the wrist joint angles in $\boldsymbol{\theta}_t^b$ to combine the outputs of $\mathcal{F}^b$ and $\mathcal{F}^h$. Appendix A.5 details the unifying process of body and hand. The resulting unified 3D human representation at time t is:

$$\mathcal{H}_t = \{\boldsymbol{\beta}_t, \tilde{\boldsymbol{\theta}}_t^b, \boldsymbol{\phi}_t^b, \boldsymbol{\tau}_t^b, \boldsymbol{\theta}_t^h\}, \tag{4}$$

with adjusted body pose $\tilde{\boldsymbol{\theta}}_t^b$. Ideally, the reconstructed SMPL-X human model at the first frame $\mathcal{H}_{t=1}$ should match the initial SMPL-X model $\mathcal{H}$ in the input scene $\mathcal{S}$. We therefore apply a global rigid transformation to match the position and orientation of $\mathcal{H}_{t=1}$ to the initial state. This transformation is analytically derived from their relative pose, and applied to all subsequent frames, resulting in:

$$\tilde{\mathcal{H}}_t = \{\boldsymbol{\beta}_t, \tilde{\boldsymbol{\theta}}_t^b, \tilde{\boldsymbol{\phi}}_t^b, \tilde{\boldsymbol{\tau}}_t^b, \boldsymbol{\theta}_t^h\}, \tag{5}$$

where $\tilde{\boldsymbol{\phi}}_t^b$ and $\tilde{\boldsymbol{\tau}}_t^b$ are the adjusted global orientation and translation.

Refining 3D Human Reference via Visual HOI Alignment. As the coarse SMPL-X reconstructed in the previous stage is obtained by unifying the outputs of two independent estimators, it is not perfectly aligned with the reference video $\mathcal{V}$ or the 3D object $\mathcal{O}$ in the scene $\mathcal{S}$, as shown in Fig. 6.

To address this, we present Visual HOI Alignment, which optimizes the body and hand pose parameters to align with both the reference video $\mathcal{V}$ and the initial 3D object state $\mathcal{O}$. We optimize body pose parameters $\{\hat{\boldsymbol{\theta}}_t^b\}_{t=1}^F$ and hand parameters $\{\hat{\boldsymbol{\theta}}_t^h\}_{t=1}^F$ by minimizing the following objective:

$$\mathcal{L}_{\text{total}} = w_b\mathcal{L}_b + w_h\mathcal{L}_h + w_{tc}\mathcal{L}_{tc} + w_{\text{HOI}}\mathcal{L}_{\text{HOI}}, \tag{6}$$

where $\mathcal{L}_b$, $\mathcal{L}_h$, $\mathcal{L}_{tc}$, $\mathcal{L}_{\text{HOI}}$ are the 2D body projection loss, 2D hand projection loss, temporal consistency loss, and HOI loss, with their corresponding weights, w_b, w_h, w_{tc}, and w_{HOI}. The 2D body projection loss $\mathcal{L}_b$ and 2D hand projection loss $\mathcal{L}_h$ are defined to align the projection of SMPL-X joints to the original joint estimations from the $\mathcal{F}^b$ and $\mathcal{F}^h$ in 2D as follows:

$$\mathcal{L}_b = \sum_{t=1}^{F} \|\Pi\left(\mathcal{J}_b(\mathcal{H}_t)\right) - \boldsymbol{j}_t^{\text{body}}\|^2, \quad \mathcal{L}_h = \sum_{t=1}^{F} \|\Pi\left(\mathcal{J}_h(\mathcal{H}_t)\right) - \boldsymbol{j}_t^{\text{hand}}\|^2, \tag{7}$$

where $\mathcal{J}_b$, $\mathcal{J}_h$ are the body and hand joint regressors from SMPL-X parameters in $\mathcal{H}_t$. The $\boldsymbol{j}_t^{\text{body}}$ and $\boldsymbol{j}_t^{\text{hand}}$ are estimated 2D body and hand joints obtained from F_b and F_h. The temporal consistency loss $\mathcal{L}_{tc}$ encourages temporal consistency, and is defined as:

$$\mathcal{L}_{tc} = \sum_{t=1}^{F-1} \mathcal{D}_{\text{geo}}(\boldsymbol{\theta}_t^b, \boldsymbol{\theta}_{t+1}^b) + \sum_{t=1}^{F-1} \mathcal{D}_{\text{geo}}(\boldsymbol{\theta}_t^h, \boldsymbol{\theta}_{t+1}^h), \tag{8}$$

where $\mathcal{D}_{\text{geo}}(\cdot)$ is a mean geodesic distance between rotations. Additionally, we present an HOI loss $\mathcal{L}_{\text{HOI}}$ to enforce the human body parts to be in contact with the object at a certain time instance:

$$\mathcal{L}_{\text{HOI}} = \min_t \mathcal{D}_{\text{chamfer}}(\mathcal{J}_*(v_t^{\text{SMPLX}}), v_*) \tag{9}$$

where $\mathcal{D}_{\text{chamfer}}(A, B)$ is a one-sided Chamfer (Fan et al., 2017) distance from A to B, $\mathcal{J}_*$ is the SMPL-X joint regressor for specific body parts (*e.g.*, left hand), v_t^{SMPLX} is the SMPL-X vertices at time t, and $v_* \in \mathbb{R}^{n_*\times3}$ is the initialized object vertices. Note that the pair $(\mathcal{J}_*, v_*)$ is specified in the text prompt we used for generating video (*e.g.*, "holds the coke with left hand"). The HOI loss is motivated by the intuition that at least one frame of the human motion should establish contact with the initial object to move the object. The resulting SMPL-X joint positions and rotations after our Visual HOI Alignment construct our 3D human reference $\hat{h}$.

Generating 2D Object Reference. For the object side, we first identify the visible vertices and their corresponding projections for the object mesh via ray casting (Roth, 1982) using the camera Π. The projected vertices are tracked across video frames using a video tracker (Karaev et al., 2024), constructing our 2D object reference $\hat{\boldsymbol{o}} = \{\hat{\boldsymbol{x}}_t\}_{t=1}^F$. We filter out vertices that are heavily occluded across video frames using the occlusion mask estimated by the tracker.

3.3 Learning Humanoid Control Policy

We formulate physics-based character control as a Markov Decision Process and optimize the policy with PPO (Schulman et al., 2017). Following MaskedMimic (Tessler et al., 2024), the human state $\boldsymbol{s}_t^h \in \mathbb{R}^{778}$ consists of joint positions, rotations, and linear/angular velocities. The object state $\boldsymbol{s}_t^o \in \mathbb{R}^{15}$ includes its position, orientation, and velocities. The action $\boldsymbol{a}_t \in \mathbb{R}^{51\times 3}$ defines PD target for all body and hand joints. To train humanoid control policy $\pi_\theta(\boldsymbol{a}_t|\boldsymbol{s}_t, \boldsymbol{g}_t)$, we define $\boldsymbol{g}_t = \hat{\boldsymbol{h}}_t^{t+k}$, the future k entries of $\hat{\boldsymbol{h}}$ as a goal, while guiding to match $\hat{\boldsymbol{o}}$ using our hybrid tracking reward.

Hybrid Tracking Reward. Our hybrid tracking reward tracks the imitation targets from the video:

$$R(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{g}_t) = R_h \cdot R_o \cdot R_{\text{contact}}, \tag{10}$$

where R_h, R_o, and R_{contact} are the human tracking, object tracking, and contact rewards. The human tracking reward R_h encourages the humanoid character to track the synthetic 3D human motion $\hat{\boldsymbol{h}}$. It is defined as the product of several joint difference rewards and a power penalty as follows:

$$R_h = r_{\text{jp}} \cdot r_{\text{jv}} \cdot r_{\text{jr}} \cdot r_{\text{lp}}^h \cdot r_{\text{lr}}^h \cdot r_{\text{pw}}, \tag{11}$$

where r_{jp}, r_{jv}, r_{jr} are the full body joint position, velocity, and rotation rewards. The terms r_{lp}^h and r_{lr}^h are the local hand joint position and rotation reward. These local joint quantities are represented in the wrist-centric coordinate by transforming joint positions and rotations relative to the wrist. The r_{pw} is a power penalty reward (Lee et al., 2023) preventing excessive forces and guiding smooth motion. Our rewards are formulated as an exponential of the mean distance between the reference and the simulated state. The object tracking reward R_o drives the 2D projection of the simulated object toward the reference trajectory $\hat{\boldsymbol{o}} = \{\hat{\boldsymbol{x}}_t\}_{t=1}^F$, defined as:

$$R_o = e^{\lambda_o \|\hat{\boldsymbol{x}}_t - \boldsymbol{x}_t\|^2}, \tag{12}$$

where t denotes the simulation time step, λ_o is a weight coefficient, and $\boldsymbol{x}_t \in \mathbb{R}^{M\times 2}$ is the projection of the same M vertices of the simulated object into the image space via the same view with Π. In practice, the object starts from its initial 3D state $\mathcal{O}$ in simulation and moves only through physical contact, guided by our object reward. We detail our reward including coefficients and formulation with contact reward in Appendix A.9 and A.10, and network architecture in Appendix A.8.

4 Experiments

4.1 Experimental Setup

Imitating Generated Videos. We evaluate DeVI on 10 generated videos, each containing a different object and a different interaction. As no prior physics-based character control method imitates dexterous HOI directly from a generated video, we compare our imitation targets rather than complete systems. Specifically, we replace our 2D object tracks with the 3D object motion recovered by RIGVid (Patel et al., 2026), DAViD (Kim et al., 2025), and Dex4D (Kuang et al., 2026) from the same video, using the corresponding object tracking rewards. All settings share the same human reference refined by Visual HOI Alignment and all other rewards, with object rewards and termination criteria matched to each object reference. As the generated video provides no ground-truth 3D motion, we treat the reconstructed 3D human motion as a pseudo ground truth and report MPJPE (Mean Per Joint Position Error) for body, hand, and all joints. For the object side, we compare the mask of the simulated object against the video mask obtained from SAM 3.1 (Carion et al., 2025), reporting pixel-level IoU, precision, recall, and F1 score.

Imitating MoCap Demonstrations. While our goal is to imitate synthetic videos, we additionally compare with methods that imitate 3D MoCap HOI demonstrations in physics simulators. Specifically, we compare DeVI against state-of-the-art 3D HOI imitation studies, PhysHOI (Wang et al., 2023), SkillMimic (Wang et al., 2025a), and InterMimic (Xu et al., 2025), on the GRAB (Taheri et al., 2020) dataset. Since DeVI uses 2D object trajectories rather than the 6D poses used by the baselines, we project 1024 vertices sampled from the object surface through a virtual camera and use their per-frame projections as the imitation target. We train all baselines under the same single-motion setting on 30 HOI motions from the GRAB dataset, and report human MPJPE separately for body, hand, and all joints along with the root joint error T_{root}. For the object side, we additionally follow CHOIS (Li et al., 2024) and report metrics for the object translation T_{obj} and orientation O_{obj}.

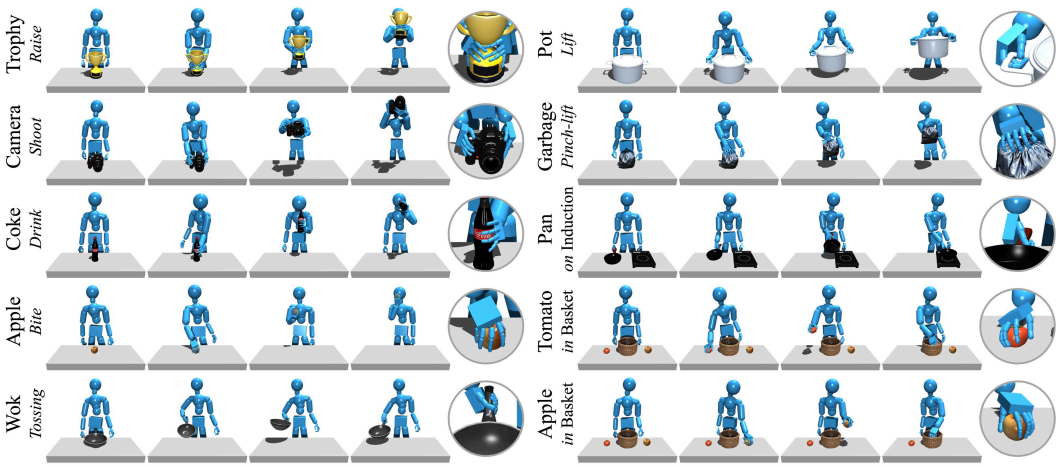


Figure 3: **Qualitative Results on Various Objects.** DeVI leverages a video diffusion model as an HOI-aware motion planner, allowing simulation of HOI with diverse objects through text prompts.

Methods	MPJPE (mm) ↓			IoU (%) ↑	Precision (%) ↑	Recall (%) ↑	F1 (%) ↑	Success (%) ↑	
	Body	Hand	All					Frame	Clip
RIGVid	641.4	684.3	666.1	38.03	45.70	43.43	44.29	36.24	10.00
DAViD	528.9	571.8	553.7	39.39	46.95	44.00	45.12	37.00	10.00
Dex4D	663.8	756.4	717.2	40.60	47.01	46.49	46.54	29.29	10.00
DeVI	**27.23**	**29.75**	**28.69**	**68.81**	**80.85**	**77.71**	**78.36**	**83.33**	**90.00**

Table 1: **Quality of Video Imitation.** We evaluate the imitation quality on synthetic videos.

Methods	Vertex Err. ↓		Rotation Err. ↓		Position Err. (m) ↓
	2D (px)	3D (m)	2D (deg.)	3D (deg.)	
RIGVid	235.6	0.4023	54.63	83.12	0.3955
DAViD	10.40	0.2808	28.10	**47.91**	0.2763
Dex4D	19.01	**0.2277**	30.67	59.10	**0.2237**
DeVI	**6.945**	N/A	**26.49**	N/A	N/A

Table 2: **Representation Choice.** We evaluate the quality of object motion recovery.

4.2 Synthetic HOI Video Imitation

Functional Manipulation for Various Objects. By leveraging a video diffusion model as an HOI-aware motion planner, DeVI allows planning various HOIs from text prompts. A generated video shows not only how a person grasps an object but also how the person uses the object afterward, providing a motion plan for functional manipulation. As shown in Fig. 3, DeVI simulates both single-handed interactions (*e.g.*, tossing a wok) and two-handed interactions (*e.g.*, taking photos). The planner also uses the input scene to guide target-aware placement, such as placing a pan on an induction stove, without an explicit scene understanding module. With the input scene fixed, different text prompts lead to placing either an apple or a tomato in the basket, demonstrating DeVI's text controllability. Figs. 16 and 17 in the appendix detail the video-motion alignment of our policy.

Comparison with Baselines. For video imitation, we treat a clip as successful when its average all-joint MPJPE stays below 0.2m and its average mask IoU exceeds 0.5, and we apply the same criteria per frame to obtain the frame success rate. As shown in Tab. 1, DeVI outperforms the baselines on all metrics. The baselines rely on explicit 3D object motion reconstruction, which remains unreliable due to monocular depth ambiguity and leads to severe hand-object misalignment (Fig. 4). In contrast, DeVI keeps the object target in 2D, allowing policy learning in simulation to find a physically feasible 3D object trajectory consistent with the 2D tracks. DeVI also outperforms policies trained with 4D HOI references from CARI4D and HOI-PAGE, whose reconstructions exhibit inaccurate object motion or insufficient finger articulation (Appendix Figs. 8 and 9, Tab. 8).

Representation Choice. We measure the reliability of each representation on 100 GRAB (Taheri et al., 2020) sequences selected alphabetically and rendered from a fixed viewpoint, providing exact ground-truth object motion. All baselines receive the same initial 3D scene information as DeVI. We report vertex, position, and rotation errors, using mean circular difference (Woehrer, 2026) for image-space rotation error. DAViD achieves lower 3D rotation error than RIGVid's direct 6D pose estimation and Dex4D's depth lifting, highlighting the strong rotation cues from 2D tracks anchored to known geometry (Tab. 2). However, its Perspective-n-Point (PnP) reconstruction still yields a



Figure 4: **Qualitative Comparison with Baselines.** DeVI avoids reconstructing unreliable 3D object trajectories compared to baselines which show significant human-object misalignments.

Methods	MPJPE (mm) ↓			T_{root} (mm) ↓	T_{obj} (mm) ↓	O_{obj} ↓
	Body	Hand	All			
PhysHOI	93.62	120.9	109.5	85.09	79.80	0.9157
DeVI	**27.78**	**21.18**	**23.96**	**29.54**	**21.04**	**0.8855**
SkillMimic	108.2	150.4	134.4	100.4	83.80	1.130
DeVI	**34.08**	**29.56**	**31.46**	**37.62**	**24.88**	**1.023**
InterMimic	58.58	105.8	87.63	47.80	90.82	1.207
DeVI	**37.29**	**37.67**	**37.52**	**38.05**	**26.45**	**0.9173**
$\text{DeVI}_{\text{w/o 2D Reward}}$	70.76	90.16	81.97	68.12	70.35	0.8195
DeVI	**37.53**	**31.80**	**34.21**	**41.67**	**20.89**	**0.6022**

Table 3: **Quality of MoCap Imitation.** We evaluate the imitation accuracy of sequences in GRAB dataset.

Methods	Success Ratio (%) ↑	
	$O_{\text{obj}} < 1.0$	$O_{\text{obj}} < 2.0$
PhysHOI	6.7	10.0
SkillMimic	3.3	3.3
InterMimic	16.7	43.3
$\text{DeVI}_{\text{w/o 2D Reward}}$	20.0	20.0
DeVI	**46.7**	**63.3**

Table 4: **MoCap Success Ratio.** We evaluate the success ratio of the imitation with two different thresholds of rotation metric O_{obj} along with MPJPE (All) < 0.1 and $T_{\text{obj}} < 0.1$.

position error of 0.2763 m, too large for dexterous manipulation. DeVI therefore keeps the tracks in 2D, preserving these rotation cues without relying on reconstructed depth.

Applications. Using generated videos as demonstrations, DeVI extends to egocentric settings, enabling articulated object manipulation with a simulated Inspire Hand RH56DFTP (Fig. 5). We further use hybrid targets from generated egocentric videos to optimize trajectories for real-world manipulation with a robotic arm and dexterous hand (Appendix B.4).

4.3 HOI MoCap Imitation

Imitation Accuracy. As the baselines use different early-termination conditions, the tracking errors of failed rollouts start showing poor accuracy at different frames and are not fairly comparable across methods. We therefore retain motions where both DeVI and the paired baseline does not collapse. Specifically, we compare motions with MPJPE (All) < 0.2 m and $T_{\text{obj}} < 0.2$ m to measure the general imitation accuracy with largely overlapped sequences. As shown in Tab. 3, DeVI outperforms each baseline on all metrics, demonstrating the benefits of our reward setup even when accurate 3D references are available. Notably, DeVI with our 2D tracking reward also outperforms a variant trained with the conventional 6D pose tracking reward, showing that the 2D trajectory, which is easier to obtain than 6D pose, can lead to stronger HOI imitation performance. We attribute this to our implicit 2D supervision, which guides object position and rotation jointly through projected geometry and avoids over-constraining the object trajectory during policy learning.

Imitation Success. To evaluate precise imitation for dexterous HOI, we define success with two different thresholds, which share the conditions MPJPE (All) < 0.1 m and $T_{\text{obj}} < 0.1$ m but differ in object orientation O_{obj}. As shown in Tab. 4, DeVI achieves higher success rates, demonstrating that hybrid targets make feasible interactions easier to learn than conventional 6D pose targets.

Methods	MPJPE (px) ↓			IoU (%) ↑	Precision (%) ↑	Recall (%) ↑	F1 (%) ↑	Success (%) ↑	
	Body	Hand	All					Frame	Clip
DeVI$_{\text{w/o VHA}}$	94.04	90.19	91.82	39.76	48.65	45.62	46.62	40.57	10.00
DeVI	**34.21**	**11.86**	**21.32**	**68.81**	**80.85**	**77.71**	**78.36**	**83.33**	**90.00**

Table 5: **Quantitative Ablation on Visual HOI Alignment.** We compare the policy learning results of DeVI with and without Visual HOI Alignment. The correction of hand-object misalignment significantly improves both the motion quality and imitation success rates of the final policy.

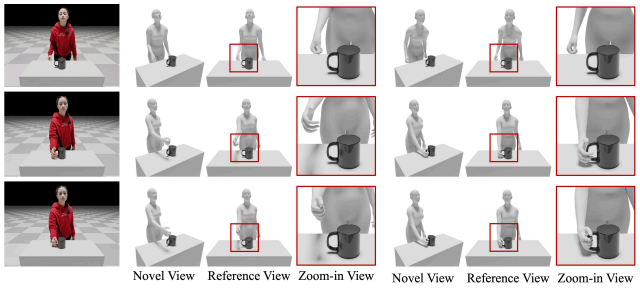


Figure 6: **Qualitative Ablation on Visual HOI Alignment.** Visual HOI Alignment corrects reconstructed human motion to match the video and the initial 3D object configuration.

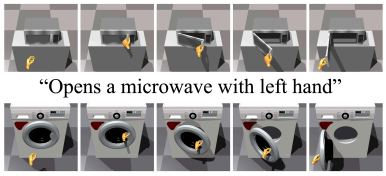


Figure 5: **Articulated Object Manipulation.** DeVI can be extended to articulated object manipulation using egocentric video generation.

Methods	MPJPE (mm) ↓		
	Body	Hand	All
TRAM	387.9	–	–
PromptHMR	75.2	94.2	85.9
PromptHMR + HaMeR	75.2	101.4	90.3
GVHMR	55.7	88.6	75.2
GVHMR + HaMeR	55.7	89.9	76.0
DeVI	**53.9**	**69.4**	**63.1**

Table 6: **Comparison with HMR.** We evaluate the quality of human mesh recovery especially in HOI manipulation scenario, showing the efficacy of Visual HOI Alignment.

4.4 Visual HOI Alignment

Ablation Study. To verify the efficacy of our Visual HOI Alignment, we conduct qualitative and quantitative ablations. As shown in Fig. 6, merging independently estimated body and hand poses into a single SMPL-X (Pavlakos et al., 2019) model leaves severe misalignment with the video and the 3D object, especially around hand-object interactions. In contrast, our Visual HOI Alignment aligns the motion with both the video and the 3D object, making the reference feasible for physical interaction. This improvement is also reflected in Tab. 5, where Visual HOI Alignment substantially improves imitation accuracy and success rates. Notably, while the comparison in Tab. 1 uses a shared 3D human reference, this ablation uses different human references. We therefore measure image-space MPJPE between 2D keypoints extracted from the video and the projected simulated joints.

Comparison with Human Mesh Recovery. For a quantitative evaluation, we compare our reconstructed human motion with existing human mesh recovery methods on the grab sequences of Subject 1 in ARCTIC (Fan et al., 2023), which provides RGB images and ground-truth human motion. As shown in Tab. 6, DeVI with Visual HOI Alignment outperforms baselines for all metrics. Notably, merging independent body and hand estimates alone does not improve hand accuracy, whereas Visual HOI Alignment substantially improves it.

5 Discussion

We present DeVI, which learns physics-based character control for dexterous HOI from text-conditioned synthetic videos without requiring 3D HOI MoCap demonstrations. Inaccurate object motion estimates and hand-object misalignment make reliable imitation targets difficult to obtain. To address this, our hybrid imitation targets bypass explicit 3D object motion reconstruction, allowing policy learning to find physically feasible object trajectories consistent with the 2D tracks. Visual HOI Alignment further refines the human reference to match the video and initial 3D object configuration, improving reconstruction quality and imitation success. Experiments demonstrate the benefits of hybrid tracking in both video and MoCap imitation. Its effectiveness even with accurate 3D references suggests advantages beyond avoiding reconstruction noise. Our functional manipulation results and egocentric applications demonstrate how generated videos can serve as motion plans for dexterous physical interaction in simulation and the real world.

AI USE STATEMENT

In this work, we used generative AI tools for generate synthetic data sets, assist with translation, design or provide feedback on research methodology or experiments, and implement methods. We have not used generative AI tools for help develop theoretical models or conceptual frameworks, assist in the writing of proofs, propose or refine hypotheses, clean and reformat dataset, support qualitative and thematic data analysis, interpret results, formulate mathematical claims, provide critical ingredients for proving mathematical claims are not applicable to this work. Additionally, we used generative AI tools for suggest experimental parameters. We have reviewed all AI-assisted work. We checked LLM-generated research ideas for potential plagiarism through a manual literature survey. LLM-generated code was verified and tested for correctness by the author. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To support reproducibility, we describe our pipeline in Sec. 3 and the datasets, baseline settings, and evaluation metrics in Sec. 4. Appendix A details scene initialization and rendering, video generation prompts, human reconstruction and alignment, contact annotation, network architecture, and reward definitions. For policy learning, Appendix A.11 specifies rollout initialization, early termination, optimization hyperparameters, hardware, and training time. We additionally document the application settings and stage-wise evaluation protocols in Appendices B.4 and B.5, respectively. We plan to release our code and results publicly.

REFERENCES

Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3d semantic scene completion. In *CVPR*, 2022.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts. In *arXiv:2511.16719*, 2025.

Boyuan Chen, Tianyuan Zhang, Haoran Geng, Kiwhan Song, Caiyi Zhang, Peihao Li, William T. Freeman, Jitendra Malik, Pieter Abbeel, Russ Tedrake, Vincent Sitzmann, and Yilun Du. Large video planner enables generalizable robot control. In *arXiv:2512.15840*, 2025a.

Jiayi Chen, Yubin Ke, Lin Peng, and He Wang. Dexonomy: Synthesizing all dexterous grasp types in a grasp taxonomy. In *RSS*, 2025b.

Jieming Cui, Tengyu Liu, Nian Liu, Yaodong Yang, Yixin Zhu, and Siyuan Huang. Anyskill: Learning open-vocabulary physical skill for interactive agents. In *CVPR*, 2024.

Jieming Cui, Tengyu Liu, Ziyu Meng, Jiale Yu, Ran Song, Wei Zhang, Yixin Zhu, and Siyuan Huang. Grove: A generalized reward for learning open-vocabulary physical skill. In *CVPR*, 2025.

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. In *NeurIPS*, 2023.

Sai Kumar Dwivedi, Dimitrije Antić, Shashank Tripathi, Omid Taheri, Cordelia Schmid, Michael J. Black, and Dimitrios Tzionas. Interactvlm: 3d interaction reasoning from 2D foundational models. In *CVPR*, 2025.

Haoqiang Fan, Hao Su, and Leonidas Guibas. A point set generation network for 3d object reconstruction from a single image. In *CVPR*, 2017.

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *CVPR*, 2023.

Yao Feng, Hengkai Tan, Xinyi Mao, Chendong Xiang, Guodong Liu, Shuhe Huang, Hang Su, and Jun Zhu. Vidar: Embodied video diffusion model for generalist manipulation. In *arXiv:2507.12898*, 2025.

Jiawei Gao, Ziqin Wang, Zeqi Xiao, Jingbo Wang, Tai Wang, Jinkun Cao, Xiaolin Hu, Si Liu, Jifeng Dai, and Jiangmiao Pang. Coohoi: Learning cooperative human-object interaction with manipulated object dynamics. In *NeurIPS*, 2024.

Tuomas Haarnoja, Ben Moran, Guy Lever, Sandy H. Huang, Dhruva Tirumala, Jan Humplik, Markus Wulfmeier, Saran Tunyasuvunakool, Noah Y. Siegel, Roland Hafner, Michael Bloesch, Kristian Hartikainen, Arunkumar Byravan, Leonard Hasenclever, Yuval Tassa, Fereshteh Sadeghi, Nathan Batchelor, Federico Casarini, Stefano Saliceti, Charles Game, Neil Sreendra, Kushal Patel, Marlon Gwira, Andrea Huber, Nicole Hurley, Francesco Nori, Raia Hadsell, and Nicolas Heess. Learning agile soccer skills for a bipedal robot with deep reinforcement learning. In *Science Robotics*, 2024.

Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. Synthesizing physical character-scene interactions. In *Proc. ACM SIGGRAPH*, 2023.

Tairan He, Jiawei Gao, Wenli Xiao, Yuanhang Zhang, Zi Wang, Jiashun Wang, Zhengyi Luo, Guanqi He, Nikhil Sobanbabu, Chaoyi Pan, Zeji Yi, Guannan Qu, Kris Kitani, Jessica Hodgins, Linxi "Jim" Fan, Yuke Zhu, Changliu Liu, and Guanya Shi. Asap: Aligning simulation and real-world physics for learning agile humanoid whole-body skills. In *arXiv:2502.01143*, 2025.

Seokpyo Hong, Daseong Han, Kyungmin Cho, Joseph S. Shin, and Junyong Noh. Physics-based full-body soccer motion control for dribbling and shooting. In *ACM TOG*, 2019.

Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. In *ICML*, 2025.

Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018.

Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In *arXiv:2410.11831*, 2024.

Hyeonwoo Kim, Sangwon Baik, and Hanbyul Joo. DAViD: Modeling dynamic affordance of 3D objects using pre-trained video diffusion models. In *ICCV*, 2025.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *arXiv:1412.6980*, 2014.

Yuxuan Kuang, Sungjae Park, Katerina Fragkiadaki, and Shubham Tulsiani. Dex4d: Task-agnostic point track policy for sim-to-real dexterous manipulation. In *arXiv:2602.15828*, 2026.

Inhee Lee, Sangwon Baik, Sungjoo Kim, Hyeonwoo Kim, Hyunsoo Cha, and Hanbyul Joo. Simuscene: Simulation-ready compositional 3d scene reconstruction from a single image. In *arXiv:2606.03994*, 2026.

Sunmin Lee, Sebastian Starke, Yuting Ye, Jungdam Won, and Alexander Winkler. Questenvsim: Environment-aware simulated motion tracking from sparse sensors. In *Proc. ACM SIGGRAPH*, 2023.

Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. In *ACM TOG*, 2023.

Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C. Karen Liu. Controllable human-object interaction synthesis. In *ECCV*, 2024.

Kailin Li, Puhao Li, Tengyu Liu, Yuyang Li, and Siyuan Huang. Maniptrans: Efficient dexterous bimanual manipulation transfer via residual learning. In *CVPR*, 2025.

Lei Li and Angela Dai. Hoi-page: Zero-shot human-object interaction generation with part affordance guidance. In *ICML*, 2026.

Yuwei Li, Longwen Zhang, Zesong Qiu, Yingwenqi Jiang, Nianyi Li, Yuexin Ma, Yuyao Zhang, Lan Xu, and Jingyi Yu. Nimble: A non-rigid hand model with bones and muscles. In *Proc. ACM SIGGRAPH*, 2022.

Junbang Liang, Ruoshi Liu, Ege Ozguroglu, Sruthi Sudhakar, Achal Dave, Pavel Tokmakov, Shuran Song, and Carl Vondrick. Dreamitate: Real-world visuomotor policy learning via video generation. In *CoRL*, 2024.

LightX2V Contributors. Lightx2v: Light video generation inference framework. https://github.com/ModelTC/lightx2v, 2025.

Yuhang Lin, Yijia Xie, Jiahong Xie, Yuehao Huang, Ruoyu Wang, Jiajun Lv, Yukai Ma, and Xingxing Zuo. Simgenhoi: Physically realistic whole-body humanoid-object interaction via generative modeling and reinforcement learning. In *arXiv:2508.14120*, 2025.

Libin Liu and Jessica Hodgins. Learning basketball dribbling skills using trajectory optimization and deep reinforcement learning. In *ACM TOG*, 2018.

Yuke Lou, Yiming Wang, Zhen Wu, Rui Zhao, Wenjia Wang, Mingyi Shi, and Taku Komura. Zero-shot human-object interaction synthesis with multimodal priors. In *arXiv:2503.20118*, 2025.

Zhengyi Luo, Jinkun Cao, Alexander W. Winkler, Kris Kitani, and Weipeng Xu. Perpetual humanoid control for real-time simulated avatars. In *ICCV*, 2023.

Zhengyi Luo, Jinkun Cao, Josh Merel, Alexander Winkler, Jing Huang, Kris M. Kitani, and Weipeng Xu. Universal humanoid motion representations for physics-based control. In *ICLR*, 2024a.

Zhengyi Luo, Jiashun Wang, Kangni Liu, Haotian Zhang, Chen Tessler, Jingbo Wang, Ye Yuan, Jinkun Cao, Zihui Lin, Fengyi Wang, et al. Smplolympics: Sports environments for physically simulated humanoids. In *arXiv:2407.00187*, 2024b.

Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, and Gavriel State. Isaac gym: High performance gpu-based physics simulation for robot learning. In *NeurIPS*, 2021.

NVIDIA. Isaac Sim. URL https://github.com/isaac-sim/IsaacSim.

OpenAI. Chatgpt: Optimizing language models for dialogue. https://openai.com/blog/chatgpt/, 2022. accessed Jan 18th, 2026.

OpenAI. Gpt-6 astra. https://openai.com/index/gpt-6-astra, 2026. accessed Sep 24th, 2026.

Liang Pan, Jingbo Wang, Buzhen Huang, Junyu Zhang, Haofan Wang, Xu Tang, and Yangang Wang. Synthesizing physically plausible human motions in 3d scenes. In *3DV*, 2024.

Shivansh Patel, Shraddhaa Mohan, Hanlin Mai, Unnat Jain, Svetlana Lazebnik, and Yunzhu Li. Robotic manipulation by imitating generated videos without physical demonstrations. In *ICLR*, 2026.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *CVPR*, 2024.

Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. In *ACM TOG*, 2018.

Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. In *ACM TOG*, 2021.

Xue Bin Peng, Yunrong Guo, Lina Halper, Sergey Levine, and Sanja Fidler. Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters. In *ACM TOG*, 2022.

Scott D. Roth. Ray casting for modeling solids. In *Comput. Graph. Image Process.*, 1982.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. In *arXiv:1707.06347*, 2017.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *arXiv:1506.02438*, 2018.

Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *Proc. ACM SIGGRAPH Asia*, 2024.

SketchFab. https://sketchfab.com/. accessed Jul 20th, 2025.

Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *ECCV*, 2020.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

SAM 3D Team, Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, Aohan Lin, Jiawei Liu, Ziqi Ma, Anushka Sagar, Bowen Song, Xiaodong Wang, Jianing Yang, Bowen Zhang, Piotr Dollár, Georgia Gkioxari, Matt Feiszli, and Jitendra Malik. Sam 3d: 3dfy anything in images. In *arXiv:2511.16624*, 2025.

Chen Tessler, Yoni Kasten, Yunrong Guo, Shie Mannor, Gal Chechik, and Xue Bin Peng. Calm: Conditional adversarial latent models for directable virtual characters. In *Proc. ACM SIGGRAPH*, 2023.

Chen Tessler, Yunrong Guo, Ofir Nabati, Gal Chechik, and Xue Bin Peng. Maskedmimic: Unified physics-based character control through masked motion inpainting. *ACM TOG*, 2024.

Chen Tessler, Yifeng Jiang, Erwin Coumans, Zhengyi Luo, Gal Chechik, and Xue Bin Peng. Masked-manipulator: Versatile whole-body manipulation. In *Proc. ACM SIGGRAPH Asia*, 2025.

Guy Tevet, Sigal Raab, Setareh Cohan, Daniele Reda, Zhengyi Luo, Xue Bin Peng, Amit Haim Bermano, and Michiel van de Panne. Closd: Closing the loop between simulation and diffusion for multi-task character control. In *ICLR*, 2025.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *IROS*, 2012.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. In *arXiv:2503.20314*, 2025.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024a.

Yinhuai Wang, Jing Lin, Ailing Zeng, Zhengyi Luo, Jian Zhang, and Lei Zhang. Physhoi: Physics-based imitation of dynamic human-object interaction. In *arXiv:2312.04393*, 2023.

Yinhuai Wang, Qihan Zhao, Runyi Yu, Hok Wai Tsui, Ailing Zeng, Jing Lin, Zhengyi Luo, Jiwen Yu, Xiu Li, Qifeng Chen, Jian Zhang, Lei Zhang, and Ping Tan. Skillmimic: Learning basketball interaction skills from demonstrations. In *CVPR*, 2025a.

Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. In *arXiv:2403.17346*, 2024b.

Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael J. Black, and Muhammed Kocabas. Prompthmr: Promptable human mesh recovery. In *CVPR*, 2025b.

Youpeng Wen, Junfan Lin, Yi Zhu, Jianhua Han, Hang Xu, Shen Zhao, and Xiaodan Liang. Vidman: Exploiting implicit dynamics from video diffusion model for effective robot manipulation. In *NeurIPS*, 2024.

Maximilian Woehrer. Image rotation angle estimation: Comparing circular-aware methods. In *arXiv:2603.25351*, 2026.

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 *ICLR*, 2024.

Zhen Wu, Jiaman Li, Pei Xu, and C. Karen Liu. Human-object interaction from human-level instructions. In *ICCV*, 2025.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *CVPR*, 2025.

Zeqi Xiao, Tai Wang, Jingbo Wang, Jinkun Cao, Wenwei Zhang, Bo Dai, Dahua Lin, and Jiangmiao Pang. Unified human-scene interaction via prompted chain-of-contacts. In *ICLR*, 2024.

Tianyi Xie, Haotian Zhang, Jinhyung Park, Zi Wang, Bowen Wen, Jiefeng Li, Xueting Li, Qingwei Ben, Haoyang Weng, Yufei Ye, David Minor, Tingwu Wang, Chenfanfu Jiang, Sanja Fidler, Jan Kautz, Linxi Fan, Yuke Zhu, Zhengyi Luo, Umar Iqbal, and Ye Yuan. Grail: Generating humanoid loco-manipulation from 3d assets and video priors. In *CoRL*, 2026a.

Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *ECCV*, 2022.

Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Visibility aware human-object interaction tracking from single rgb camera. In *CVPR*, 2023.

Xianghui Xie, Jan Eric Lenssen, and Gerard Pons-Moll. Intertrack: Tracking human object interaction without object templates. In *3DV*, 2025.

Xianghui Xie, Bowen Wen, Yan Chang, Hesam Rabeti, Jiefeng Li, Ye Yuan, Gerard Pons-Moll, and Stan Birchfield. Cari4d: Category agnostic 4d reconstruction of human-object interaction. In *CVPR*, 2026b.

Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. In *arXiv:2404.07191*, 2024.

Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, and Liang-Yan Gui. Intermimic: Towards universal whole-body control for physics-based human-object interactions. In *CVPR*, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024a.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *arXiv:2408.06072*, 2024b.

Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *CVPR*, 2021.

Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *ICCV*, 2023.

Jason Y. Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *ECCV*, 2020.

Youliang Zhang, Ronghui Li, Yachao Zhang, Liang Pan, Jingbo Wang, Yebin Liu, and Xiu Li. A plug-and-play physical motion restoration approach for in-the-wild high-difficulty motions. In *ICCV*, 2025a.

Ziyu Zhang, Sergey Bashkirov, Dun Yang, Yi Shi, Michael Taylor, and Xue Bin Peng. Physics-based motion imitation with adversarial differential discriminators. In *Proc. ACM SIGGRAPH Asia*, 2025b.

Siheng Zhao, Yanjie Ze, Yue Wang, C. Karen Liu, Pieter Abbeel, Guanya Shi, and Rocky Duan. Resmimic: From general motion tracking to humanoid whole-body loco-manipulation via residual learning. In *arXiv:2510.05070*, 2025.

A Implementation Details

A.1 Scene Initialization

Our scene initialization follows a tabletop scenario. We place an SMPL-X (Pavlakos et al., 2019) human at the origin on the xy-plane and initialize it to face the positive y-axis. We then place a table on the xy-plane at $(x, y) = (0.0, 0.4)$, with dimensions $(1.0, 0.5, 0.8)$. The objects are initialized on the table to construct our scene. We assume a physically valid state (*e.g.*, no floating objects) for the initialized scene and assign an initial pose to each object accordingly, using the same poses to initialize the physics simulation.

A.2 Deforming Textured Human

We replace the SMPL-X model with a 3D textured human mesh by deforming the mesh to match the pose of the initialized SMPL-X. We use the 3D textured human meshes from the THuman 2.0 dataset (Yu et al., 2021) with their corresponding SMPL-X annotations to assign SMPL-X offsets and skinning weights to each vertex of the textured human mesh. For the textured human mesh vertices $\{x_i\}_{i=1}^{N} \in \mathbb{R}^{N\times 3}$ and $\{v_j\}_{j=1}^{M} \in \mathbb{R}^{10475\times 3}$, we assign the offset $o_i \in \mathbb{R}^3$ and skinning weight $w_i \in \mathbb{R}^J$ to the vertex x_i as follows:

$$w_i = \sum_{k\in\mathcal{N}_K(i)} \alpha_{i,k} W_k, \quad o_i = \sum_{k\in\mathcal{N}_K(i)} \alpha_{i,k} O_k,$$

where $\mathcal{N}_K(i)$ is the set of indices of the K-nearest vertices of x_i in $\{v_j\}_{j=1}^{M}$, $W_k \in \mathbb{R}^J$ is the skinning-weight vector for the $J = 55$ joints at SMPL-X vertex v_k, $O_k \in \mathbb{R}^3$ is the offset of the SMPL-X vertex v_k computed via shape and pose parameters, and $\alpha_{i,k}$ is the blending coefficient of each skinning weight. The coefficient $\alpha_{i,k}$ is defined using a Gaussian kernel over the distance between x_i and its nearest neighbors $\{v_k\}_{k\in\mathcal{N}_K(i)}$ as follows:

$$\alpha_{i,k} = \frac{e^{s_{i,k}}}{\sum_{k'\in\mathcal{N}_K(i)} e^{s_{i,k'}}}, \quad s_{i,k} = -\frac{\|x_i - v_k\|^2}{2\sigma^2}$$

Using the assigned offset o_i and skinning weight w_i, we deform the 3D textured human mesh via linear blend skinning. In practice, we use $K = 16$ for approximating the offsets and skinning weights.

A.3 Rendering Scene

For the initialized scene, we initially place 16 candidate cameras at a radius of 1.5, spaced at $45°$ azimuth intervals, and facing the origin with two elevation angle variations, $15°$ and $30°$. The height of the cameras is adjusted about 1.0 along the positive z-axis to capture the scene within the image frame. Among the 16 cameras, we empirically find that selecting one of the six frontal views in which both the human's hands and the object are visible is beneficial, especially for hand pose estimation. For the remaining intrinsic parameters, we follow Isaac Gym (Makoviychuk et al., 2021), and define the intrinsic matrix K as follows:

$$K = \begin{bmatrix} f & 0 & w/2 \\ 0 & f & h/2 \\ 0 & 0 & 1 \end{bmatrix}, \tag{13}$$

where $f = w/2$ and h, w are the height and width of the image. In practice, we use $h = 576$, $w = 1024$ for rendering. The installed camera is used to render an image, which is used as the input for generating motion plans in the form of a video using a video diffusion model.

A.4 Generating HOI Video

To generate an HOI video from the rendered images, we use a pre-trained video diffusion model, Wan (Wan et al., 2025) with additional LightX2V (LightX2V Contributors, 2025) LoRA for faster inference. Generating a single video takes about 10 minutes on an NVIDIA A6000 GPU. For the text prompt, we use the following format:

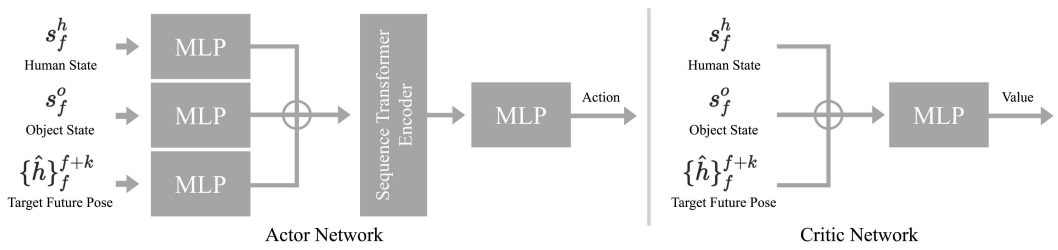


Figure 7: **Network Architecture of DeVI.** Our humanoid control policy network consists of a transformer-based actor network and an MLP-based critic network with the same input states.

```
A person [action] [object] with [hand index] [detail].  Both hands
are visible in the frame.  Stationary camera.
```

We use both hand-designed prompts and prompts automatically generated by ChatGPT (OpenAI, 2022) to fill in the brackets. While keeping the format, we additionally adjust the word "person" and gendered pronouns in the prompt according to the gender of the 3D textured human mesh, and add details about the non-interacting hand.

A.5 Unifying Body and Hand Estimators

To reconstruct the coarse human from the generated video, we use the body estimator GVHMR (Shen et al., 2024) and the hand estimator HaMeR (Pavlakos et al., 2024). To unify the outputs of each estimator into a single SMPL-X, we replace the body estimator's local wrist pose with the following transformation:

$$G_e = \prod_{r=0}^{e} R_{p_r}$$
$$R_w = G_e^{-1} G_w$$

, where $R_{p_r} \in SO(3)$ is the rotation of each body joint p_r obtained from the body estimator, and G_w is the global rotation of the hand (in this case, the wrist) obtained from the hand estimator. We define $(p_0, p_1, \ldots, p_e)$ as an ordered sequence of joints along the forward-kinematics chain from the pelvis to the elbow. We omit the left/right subscripts for simplicity. The local rotation of the wrist R_w is converted to axis-angle representation and used for the unified SMPL-X's wrist pose.

A.6 Details of Visual HOI Alignment

To reconstruct human motion aligned with both the video and the existing 3D object, our Visual HOI alignment optimizes SMPL-X body and hand parameters using a 2D projection loss and a one-sided Chamfer distance loss. Even though we compute the loss using 19 body joints and 32 hand joints, optimizing all body parameters adversely affects the estimates for occluded joints (especially in the lower body). In practice, we optimize only the upper-body poses, specifically for the hands, wrists, elbows, shoulders, and spine joints. We use the Adam (Kingma & Ba, 2014) optimizer with a learning rate of 2×10^{-2} without decay. For each loss, we use $w_b = 1.0$, $w_h = 1.0$, $w_{tc} = 1.0 \times 10^4$, and $w_{\text{HOI}} = 5.0 \times 10^2$.

A.7 Contact Estimation

Following GRAIL (Xie et al., 2026a), we annotate human-object contact from video frames using a vision-language model (VLM). For each frame, we provide the VLM (OpenAI, 2026) with the image and the target object name to obtain a contact decision and confidence level in JSON format. The prompt defines contact as physical touching by any body part, excludes proximity without touching, and instructs the model to return false when uncertain. We set the binary pseudo contact label ψ_t to 1 only when `in_contact` is `true` and `confidence` is `high`, and to 0 otherwise. These global labels provide contact timing cues, while the contact reward is evaluated only for participating

hands, as detailed in Appendix A.10. For every frame, we use the following prompt, replacing `{object_name}` with the target object name.

```
You are a careful vision classifier. Given an image of a person
and an object, determine whether the person is in physical contact
with the object.
Object: {object_name}
Guidelines:
- Contact means the person is physically touching or holding the
object.
- Near proximity without touching does NOT count as contact.
- If any part of the person's body touches the object, return true.
- Be conservative: when uncertain, return false.
Respond with strictly valid JSON using this schema (no extra text):
{"in_contact": true/false, "confidence": "high/medium/low"}
```

A.8 Network Architecture

Our humanoid control policy network uses an actor-critic architecture. Both the actor and critic networks take the human state, object state, and target future pose as inputs. Each input of the actor network first passes through a separate 2-layer MLP with 256 hidden units. The resulting features are then concatenated and fed into a sequence transformer encoder for encoding. The encoded latent passes through a 3-layer MLP with 1024 hidden units and outputs the action of the humanoid. In contrast, the three inputs of the critic network are flattened and concatenated without a separate encoding stage, and pass through a 4-layer MLP with 1024 hidden units to output the value of the given state. The overall architecture is shown in Fig. 7.

A.9 Details about Hybrid Tracking Reward

Each of our tracking term shares the same exponential form

$$r_k = \exp\big(c_k \cdot \bar{e}_k\big), \tag{14}$$

where $\bar{e}_k$ is the tracking error between the simulated state and the reference $\hat{h}$ (a mean squared distance for position terms and a mean geodesic angular distance for rotation terms), and $c_k < 0$ is a sensitivity coefficient. The error terms and coefficients used in our experiments are summarized in Table 7. For the full-body terms ($r_{\text{jp}}, r_{\text{jv}}, r_{\text{jr}}$), positions, velocities, and rotations are expressed in the global frame and averaged over all body links. For the local hand terms ($r^h_{\text{lp}}, r^h_{\text{lr}}$), finger keypoints are first translated to the wrist origin and rotated into the wrist frame. The power penalty r_{pw} follows (Lee et al., 2023) and discourages excessive joint torques, with τ_j and $\dot{q}_j$ denoting the actuator torque and velocity at the j-th joint. In addition, we redundantly apply the same exponential form to the fingertip locations ($T \subset H$) and to a small subset of key body links (K, the two wrists and two ankles) on top of r^h_{lp} and r_{jp}, with coefficients -100 and -120 respectively, to place stronger emphasis on tracking these contact-relevant points. For the 2D object tracking reward, we use $\lambda_o = -1.0 \times 10^{-3}$ as a coefficient.

For the DeVI variant without the 2D reward in Tabs. 3 and 4, we replace R_o with the following 6D object pose tracking reward.

$$R_{\text{6D}} = \exp\big(c_{\text{pos}} \|\boldsymbol{p}_t - \boldsymbol{p}_t^{\text{ref}}\|_2^2\big) \exp\big(c_{\text{rot}} \theta_t^2\big)\,, \tag{15}$$

where $\boldsymbol{p}_t$ and $\boldsymbol{p}_t^{\text{ref}}$ denote the simulated and reference object positions in meters, respectively, and θ_t is the angle of their relative rotation in radians. We use $c_{\text{pos}} = -80$ and $c_{\text{rot}} = -1$ in all 6D object tracking comparisons. For the generated-video comparisons, we use the same reward and coefficients with reconstructed object trajectories from RIGVid (Patel et al., 2026), DAViD (Kim et al., 2025), and Dex4D (Kuang et al., 2026). In the DeVI ablation, the human reference, other rewards, network architecture, training budget, initialization, and human termination criteria remain identical. Only the object target, tracking reward, and object termination criterion change, as detailed in Appendix A.11.

Term	Error $\bar{e}_k$	Coefficient c_k
r_{jp} (body joint position)	$\frac{1}{B}\sum_b \|p_b - \hat{p}_b\|_2^2$	-100
r_{jv} (body joint velocity)	$\frac{1}{B}\sum_b \frac{1}{3}\|v_b - \hat{v}_b\|_2^2$	-0.5
r_{jr} (body joint rotation)	$\frac{1}{B}\sum_b \theta(q_b, \hat{q}_b)$	-1.5
r^h_{lp} (local hand position)	$\frac{1}{H}\sum_h \frac{1}{3}\|p^w_h - \hat{p}^w_h\|_2^2$	-100
r^h_{lr} (local hand rotation)	$\frac{1}{H}\sum_h \theta(q^w_h, \hat{q}^w_h)$	-2.0
r_{pw} (power penalty)	$\sum_{j=1}^{D} \lvert\tau_j\, \dot{q}_j\rvert$	-2.0×10^{-5}

Table 7: **Reward Term Definitions.** Error definitions and coefficients for each tracking term.

A.10 Details about Contact Reward

The contact reward, $R_{\text{contact}} = R_{cf} \cdot R_{cd}$, encourages the humanoid to establish contact with the target object in the simulation. It is defined via the product of the contact force reward R_{cf} and the contact distance reward R_{cd}. While similar formulations appear in previous studies (Wang et al., 2023; Xu et al., 2025), our reward relies on contact timing cues explicitly inferred from the generated video $\mathcal{V}$. For each frame, we use the global binary contact annotation $\psi_t \in \{0, 1\}$ from the VLM (OpenAI, 2026), as described in Appendix A.7. For both contact terms, we include only the participating hands, denoted by $\mathcal{A} \subseteq \{L, R\}$, excluding the non-participating hand in single-hand manipulation. Using this shared contact label, we average the per-hand contact force rewards as follows:

$$R_{\text{cf}} = (1 - \psi_t) + \frac{\psi_t}{|\mathcal{A}|} \sum_{h \in \mathcal{A}} \Psi^h_t, \tag{16}$$

where $\Psi^h_t \in [0, 1]$ is the ratio of force sensors in hand h for which the measured force exceeds a predefined threshold as follows:

$$\Psi^h_t = \frac{1}{K_h} \sum_{j=1}^{K_h} \mathbb{I}[c^{h,j}_t > \tau_{\text{contact}}], \tag{17}$$

where K_h is the number of force sensors in hand h, $c^{h,j}_t \in \mathbb{R}$ is the contact force measured at its j-th joint at simulation step t, τ_{contact} is the contact force threshold, and $\mathbb{I}$ is an indicator function. Using the one-sided Chamfer distance d^h_t from the joints of hand h to object vertices, we average the per-hand contact distance rewards as follows:

$$R_{\text{cd}} = (1 - \psi_t) + \frac{\psi_t}{|\mathcal{A}|} \sum_{h \in \mathcal{A}} \sigma\left(-\lambda_{\text{c}} (d^h_t)^2\right), \tag{18}$$

where σ is a sigmoid function and λ_c is a weighting factor. In practice we use $\tau_{\text{contact}} = 0.1$ and $\lambda_c = 10.0$ in our experiments.

A.11 Training Details

Time Sampling for Initialization. Previous studies on imitating human motions (Tessler et al., 2024; Wang et al., 2025a) sample a random time step from the reference motion for initialization, which has a positive effect on sample efficiency during rollout. Unlike these studies, we do not have access to the object's 6D pose reference, which makes initialization at an arbitrary time step difficult. Instead, we propose a strategy that initializes at the pre-contact frame where the object's 6D pose is the same as the initial pose, with a 50% probability. This strategy comes from the observation that, in general, more trials are required near the frame where the human starts interacting with the object than in intervals that solely imitate human motion. We empirically find that this strategy boosts policy learning compared to first-frame initialization.

Early Termination. For sample efficiency during rollout, we perform early termination when the current state is significantly different from the reference imitation target. We define early termination based on the per-joint distance error for 3D body and fingertip joints, and the pixel distance error for the 2D object trajectory. Specifically, we terminate if the mean error of the 3D body joints exceeds

200 mm or if any 3D body joint error exceeds 400 mm, or if the mean 3D fingertip error exceeds 40 mm. For the pixel distance error of the 2D object trajectory, we define the threshold based on the image resolution as follows.

$$\tau_{\text{2D}} = \alpha_{\text{2D}}\sqrt{W^2 + H^2} \tag{19}$$

In practice, we use $\alpha_{2D} = 0.08$, giving approximately $\tau_{2D} = 94$ pixels for $W = 1024$ and $H = 576$. For the 6D object tracking settings, we instead terminate when the mean distance between corresponding object points exceeds 0.15 m.

$$E_{\text{3D}}(t) = \frac{1}{N}\sum_{i=1}^{N} \left\| \boldsymbol{y}_{t,i} - \boldsymbol{y}_{t,i}^{\text{ref}} \right\|_2 > 0.15 \text{ m}, \qquad N = 1024, \tag{20}$$

where $\boldsymbol{y}_{t,i}$ and $\boldsymbol{y}_{t,i}^{\text{ref}}$ are the same sampled object point transformed by the simulated and reference poses, respectively. This error captures both translation and rotation and is used for termination, while the reward follows Eq. 15.

Training Setup. For training our actor-critic network, we use the Adam optimizer with a learning rate of 2×10^{-5} for the actor network and 1×10^{-4} for the critic network. We collect data using 4096 environments in Isaac Gym, update each network's parameters after 32 rollouts, and use a batch size of 1024. For hardware, we train on NVIDIA A6000 GPUs, and it takes about 20 hours on a single GPU to imitate a 263-frame video, although the runtime varies by reference motion.

Policy Learning. To learn the humanoid control policy using our hybrid tracking reward, we build actor-critic network with transformer-based actor network and MLP-based critic network. The actor network outputs the action which is used as a control signal for the humanoid, and the critic network estimates the value function which is used to compute the advantage A_t of the current state-action pair using GAE (Schulman et al., 2018). We update the actor network using the policy gradient from PPO (Schulman et al., 2017) as follows:

$$\mathcal{L}_{\text{ppo}} = -\mathbb{E}_t[\min(r_t(\psi_{\text{actor}})A_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)A_t] \tag{21}$$

$$\mathcal{L}_{\text{bound}} = \sum_i \left(\text{ReLU}(\mu_i - 1)^2 + \text{ReLU}(-\mu_i - 1)^2\right), \tag{22}$$

where $\epsilon \in \mathbb{R}$ is a clipping constant, $\text{clip}(\cdot, a, b)$ is a clip function from a to b, $r_t(\psi)$ is the ratio of the likelihoods of the current action under the updated and old policies, and μ_i is an output of the actor network which is the estimated mean of the distribution. We softly update the actor network so that the likelihood of actions with higher advantage increases, while ensuring that the output action remains bounded in $[-1, 1]$ via the following actor loss:

$$\mathcal{L}_{\text{actor}} = \mathcal{L}_{\text{ppo}} + \mathcal{L}_{\text{bound}} \tag{23}$$

Along with the actor loss, we update the critic network to predict the estimated returns, modelling a reliable value function as follows:

$$V_{\text{clip}} = \text{clip}(V_{\psi_{\text{critic}}}, V_{\text{old}} - \epsilon, V_{\text{old}} + \epsilon) \tag{24}$$

$$\mathcal{L}_{\text{critic}} = \frac{1}{2}\mathbb{E}\left[\max\left(\left(V_{\psi_{\text{critic}}} - R_t\right)^2, \left(V_{\text{clip}} - R_t\right)^2\right)\right], \tag{25}$$

where V_{old} is an old value function, $V_{\psi_{\text{critic}}}$ is the current estimated value function, and $R_t = A_t + V_{\text{old}}$ is the target return. For the clipping constant, we use $\epsilon = 0.2$.

A.12 Penetration Issue in Rigid-Body Simulation

Penetration between the humanoid and the object in physics-based human-object interaction can arise from two distinct sources. First, physics simulators commonly use simplified collision geometries that differ from the higher-resolution visual meshes used for rendering, so the simulator may correctly resolve contact between the underlying collision geometries while the visual meshes appear to overlap. Second, rigid-body simulators (Makoviychuk et al., 2021; NVIDIA; Todorov et al., 2012) resolve contacts via iterative constraint solvers under a fixed iteration budget, which inherently leaves contact constraints imperfectly satisfied. The effect is amplified in dexterous manipulation, where many simultaneous finger-object contacts must be resolved within a single discrete timestep. State-of-the-art physics-based HOI methods (Wang et al., 2023; 2025a; Xu et al., 2025) also exhibit such residual penetration, as specifically noted by MaskedManipulator (Tessler et al., 2025). Following MaskedManipulator (Tessler et al., 2025), we adopt a power penalty on the hand degrees of freedom as shown in Tab. 7.

Methods	MPJPE (px) ↓			IoU (%) ↑	Precision (%) ↑	Recall (%) ↑	F1 (%) ↑	Success (%) ↑	
	Body	Fingers	All					Frame	Clip
CARI4D	157.66	126.29	139.57	37.49	43.41	42.70	42.88	5.05	0.00
HOI-PAGE	170.41	177.71	174.63	42.68	48.88	48.67	48.58	3.67	0.00
DeVI	**34.21**	**11.86**	**21.32**	**68.81**	**80.85**	**77.71**	**78.36**	**83.33**	**90.00**

Table 8: **Quantitative Comparison with 4D HOI Reconstruction.** We compare the imitation policies trained with references from CARI4D and HOI-PAGE against DeVI. Following the evaluation in Tab. 5, we measure human MPJPE in image space using video keypoints as pseudo ground truth.

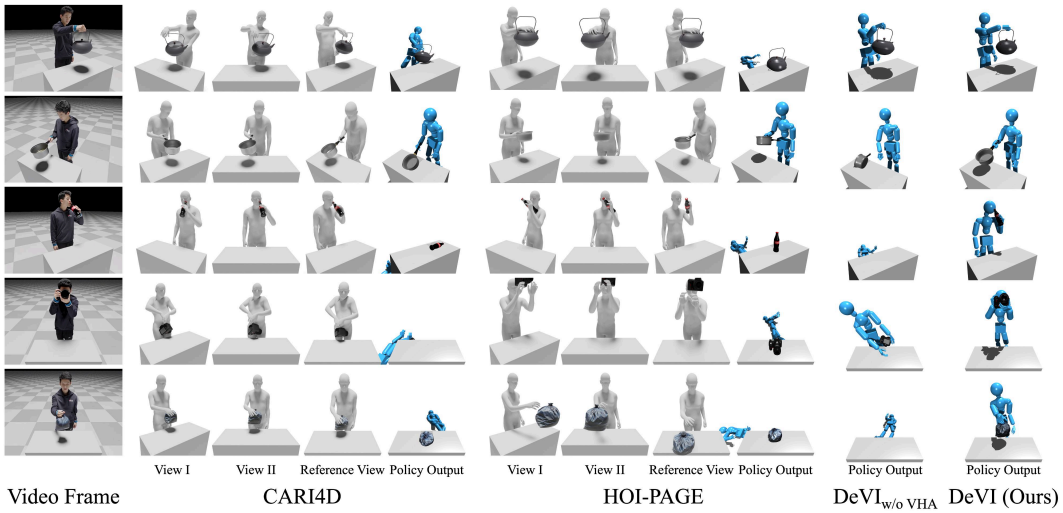


Figure 8: **Imitation from 4D HOI Reconstruction.** We compare CARI4D and HOI-PAGE with DeVI and the ablation without Visual HOI Alignment. Imprecise hand-object alignment leads to imitation failures, while DeVI successfully imitates the generated interactions.

B Additional Results

B.1 Synthetic HOI Video Imitation

Imitation from 4D HOI Reconstruction. To complement the object motion reconstruction comparison in Fig. 4, we evaluate CARI4D (Xie et al., 2026b) and HOI-PAGE (Li & Dai, 2026) using their reconstructed 4D human-object motions as imitation references. As CARI4D and HOI-PAGE each reconstruct human motion from the video, the methods use different human references. Following the evaluation in Tab. 5, we therefore measure human MPJPE in image space using 2D joints predicted from the video as ground truth. We also include DeVI without Visual HOI Alignment to evaluate the effect of correcting the human reference. As shown in Fig. 8, both reconstruction methods fail to recover the precise hand-object configurations required for the shown dexterous interactions, leading to unsuccessful policy rollouts. The ablation further shows that inaccurate human references hinder contact formation even when the object target remains in 2D. As shown in Tab. 8, DeVI outperforms both reconstruction baselines across all metrics.

Failure Analysis of 4D HOI Reconstruction. As shown in Fig. 9, HOI-PAGE produces relatively noisy object position estimates over time, disrupting alignment between the object and the hand. In the drinking sequence, the reconstructed bottle shifts before the hand establishes a grasp, providing an inconsistent reference for physical interaction. While CARI4D recovers the spatio-temporal relationship between the human and object more consistently, the reconstructed finger articulation does not form the contacts required to grasp the object. Consequently, the simulated hand reaches the bottle but fails to secure a grasp, causing the bottle to topple during imitation. Using Visual HOI Alignment and hybrid imitation targets, DeVI establishes the required contacts and effectively

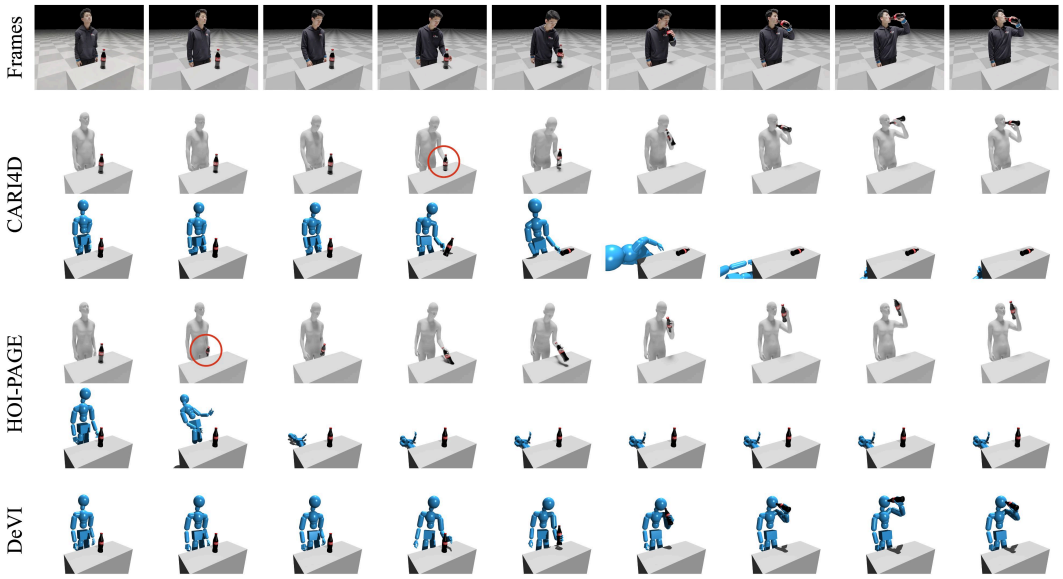


Figure 9: **Failure Modes of 4D HOI Reconstruction Baselines.** HOI-PAGE produces noisy object positions, while CARI4D lacks the finger articulation needed for grasping despite more consistent human-object motion. Both references lead to imitation failures.

imitates the generated interactions in the physics simulator without explicitly reconstructing the full 4D HOI motion.

Detailed Results with Hybrid Imitation Targets. Detailed results of the simulated HOIs, including the text prompts and intermediate hybrid imitation targets, are shown in Fig. 16 and Fig. 17. As shown in Fig. 16 and Fig. 17, the simulated HOI motions produced by policies trained to imitate the hybrid imitation targets are well aligned with the generated videos. We show that hybrid imitation targets guide the policy to track 3D human motion while discovering object poses consistent with 2D observations.

B.2 Generalizability of Universal Controllers

As high-quality 4D HOI data remain scarce, existing datasets provide limited coverage of object geometries, grasps, and manipulation patterns. While universal controllers such as InterMimic (Xu et al., 2025) imitate diverse HOI motions with a single policy, their generalizability still depends on the coverage of the training demonstrations. For this evaluation, we use InterMimic's pre-trained checkpoint to imitate precise, physically feasible 4D HOI motions obtained from simulated rollouts of policies trained with DeVI. These reference motions include both humanoid and object states over time. As shown in Fig. 10, InterMimic still fails to establish or maintain the hand-object contacts required for novel dexterous interactions (*e.g.*, drinking from a coke bottle and shooting with a camera). These results show that even precise, physically feasible 4D references do not guarantee successful imitation of novel video-based motion plans by an existing universal controller. We address this limitation by learning a separate policy for each video with our hybrid imitation targets, adapting the controller to the specific interaction.

B.3 HOI MoCap Imitation

Qualitative Comparison with Baselines. We additionally showcase the qualitative results of DeVI and the baselines in Fig. 11. Although DeVI only leverages 2D object trajectories (without full 6D poses), we show results comparable to those of the baselines. Specifically, DeVI better follows human poses compared to SkillMimic (Wang et al., 2025a) and InterMimic (Xu et al., 2025), and better follows object poses compared to PhysHOI (Wang et al., 2023). We demonstrate that our hybrid tracking reward effectively imitates MoCap data even without leveraging precise 6D object poses.

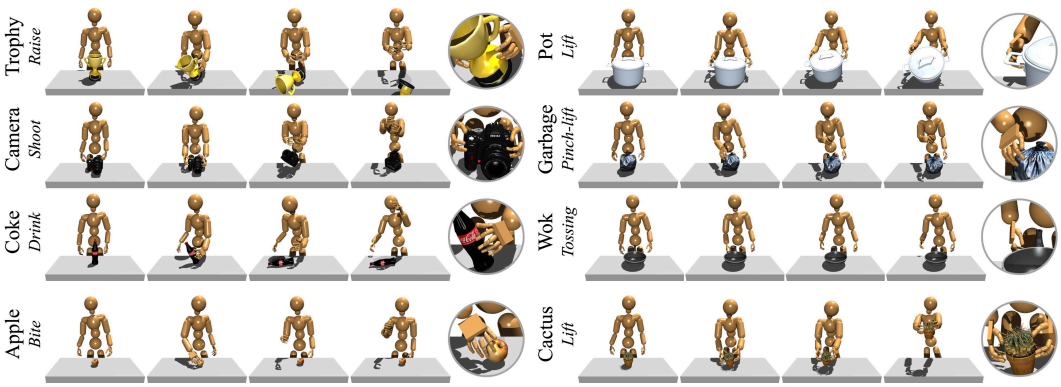


Figure 10: **Limited Generalizability of Universal Controller.** Due to the scarcity of 4D HOI data, an existing universal controller fails to imitate novel HOI motions, such as those reconstructed from videos.

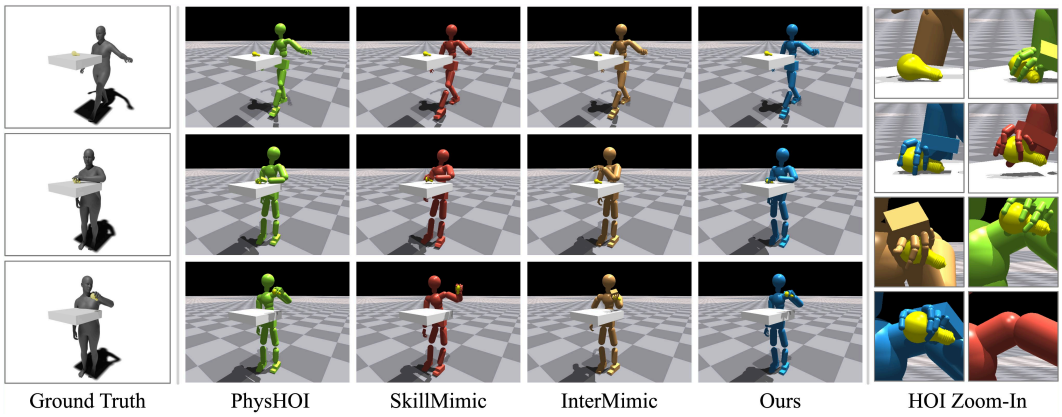


Figure 11: **Qualitative Comparison with MoCap Imitation Baselines.** Even without 6D object poses, DeVI tracks both the human and the object better than baselines using a 6D object pose reward.

DeVI on GRAB Dataset. We showcase additional qualitative results in Fig. 18 and Fig. 19. The figures show the imitation results of our humanoid control policy network on the GRAB (Taheri et al., 2020) dataset. As we use only 2D trajectories for learning the humanoid control policy, we project the 3D object vertices into a virtual camera view and use the resulting 2D trajectories as a reference imitation target. The results show that our method successfully imitates HOI motion using relatively sparse signals, even for HOI scenarios that we do not generate via a video diffusion model.

B.4 Generalization and Applications

Real-World Manipulation. While our main evaluation focuses on physics-based imitation in simulation, we additionally show the applicability of our hybrid imitation target and Visual HOI Alignment in a real-world environment using a robotic arm equipped with a dexterous hand. Specifically, we generate egocentric videos and extract hybrid imitation targets from them. Using these targets, we define the cost for model predictive control (MPC) and find a feasible robotic hand trajectory that minimizes this cost in a physics simulator. We then execute the optimized trajectory on the real robot in an open-loop manner by supplying the trajectory as PD targets. Following Dexonomy (Chen et al., 2025b), we additionally tighten the grasp at contact onset to hold the object firmly. As shown in Fig. 12, the robot performs pick-and-place motions with everyday objects. These examples demonstrate coordinated arm and finger motion for grasping, transporting, and releasing objects with different geometries in a physical environment.

Generalization beyond Controlled Environments. As DeVI leverages synthetic videos instead of 3D motion-capture data, it generalizes to novel scenes, even from a single scene image. Specifically,

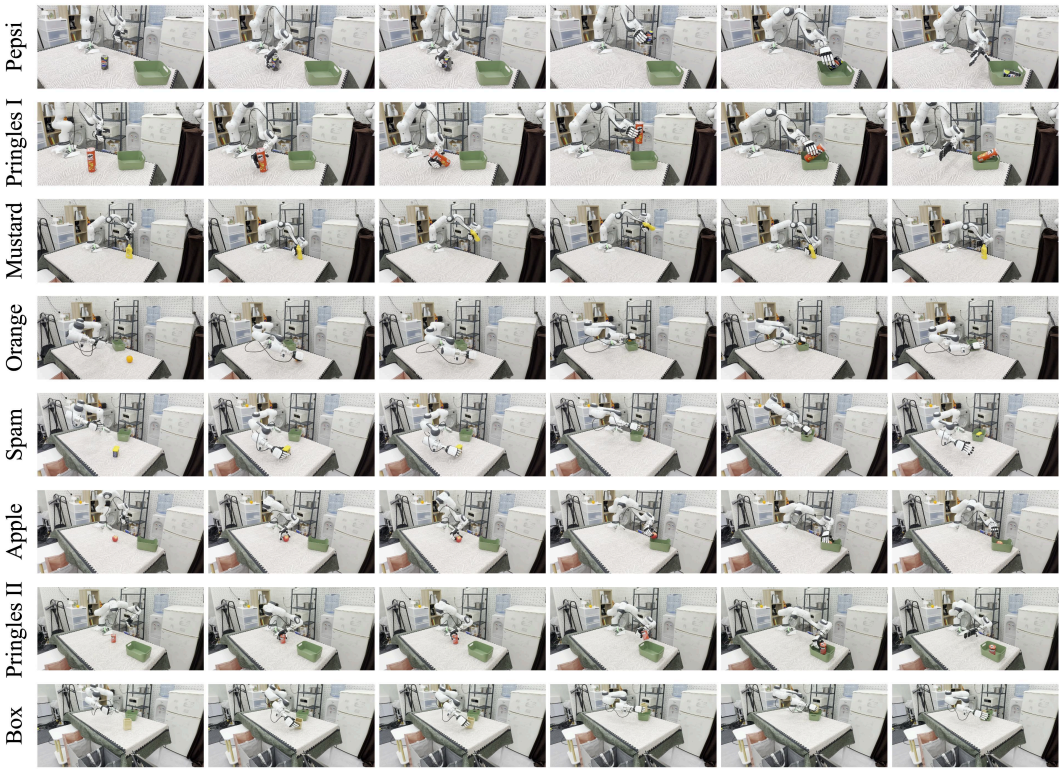


Figure 12: **Real-World Dexterous Manipulation.** Our hybrid imitation target serves as an effective cost for MPC (Model Predictive Control) to optimize a the trajectory that can be executed in the real world using a robotic arm equipped with a dexterous hand.

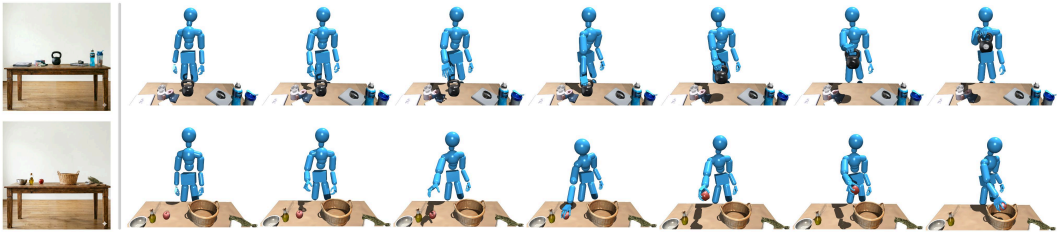
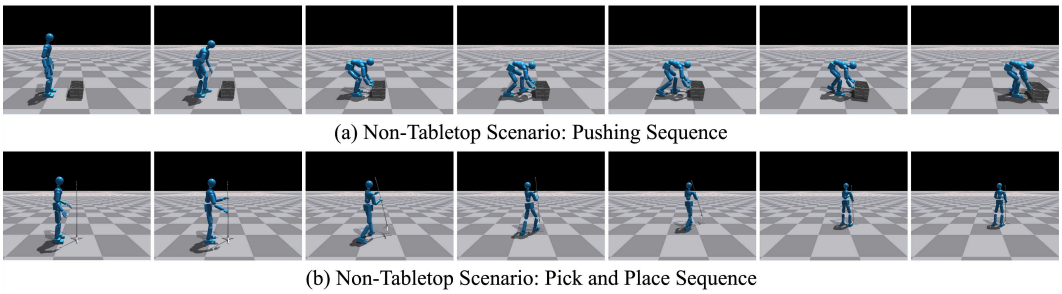


Figure 13: **Generalization of DeVI beyond controlled environments.** DeVI generalizes to novel scenes generated via Image-to-3D.

we generate a reference scene image from a text prompt using Gemini (Team et al., 2023) and use SimuScene (Lee et al., 2026) to construct the simulation-ready scene with various assets. We replace the initial scene with the reconstructed scene to generate physics-based character motion interacting with the objects in the scene. As shown in Fig. 13, HOI motions within a reconstructed scene can be generated inside a physics simulator, demonstrating our pipeline is broadly applicable across a wide range of environments.

Non-tabletop Scenarios. While we intentionally scoped our evaluation to tabletop setups to focus on dexterous HOI imitation from synthetic video without relying on 3D MoCap signals, our hybrid tracking reward is not limited to tabletop scenarios. Fig. 14 shows the results of DeVI trained on the FullBodyManip (Li et al., 2023) dataset. As shown in Fig. 14, DeVI and the hybrid imitation rewards extend beyond tabletop scenarios and can also be applied to non-tabletop environments.

Articulated Object Manipulation. For the articulated object manipulation in Fig. 5, we initialize a NIMBLE (Li et al., 2022) hand mesh in the 3D scene and render the initial image from an egocentric viewpoint. We manually frame the hand mesh so that the cut surface at the wrist lies below the



(a) Non-Tabletop Scenario: Pushing Sequence

(b) Non-Tabletop Scenario: Pick and Place Sequence

Figure 14: **Non-Tabletop Scenarios.** Our hybrid imitation targets are not limited to tabletop scenarios and can be applied to non-tabletop motions such as (a) pushing and (b) pick and place.

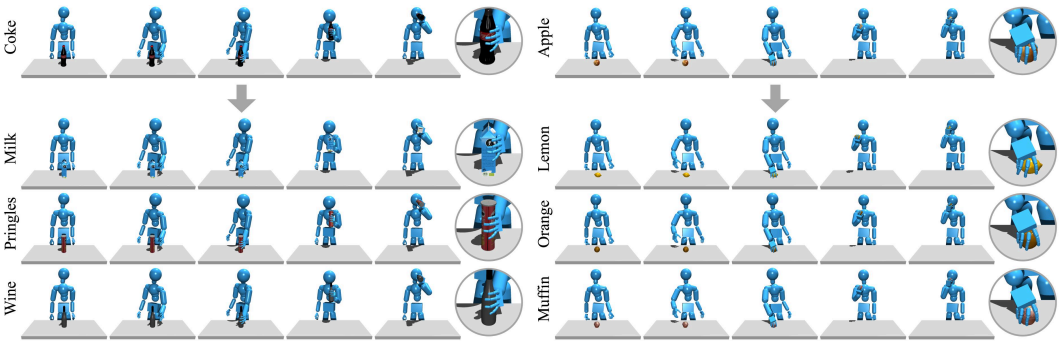


Figure 15: **Zero-Shot Cross-Object Policy Transfer.** Our policy trained to imitate the video can be transferred to an unseen object with similar geometry.

bottom image boundary, while the hand remains visible. We empirically find that this initialization helps the video generator account for the hand scale when synthesizing the interaction. Following ManipTrans (Li et al., 2025), we retarget the hand trajectory estimated from the generated video to the simulated Inspire Hand RH56DFTP, providing the hand reference for policy learning.

Cross-Object Policy Transfer. As shown in Fig. 15, our trained policy can be transferred to unseen object, allowing it to generate the same motion while manipulating different objects. Specifically, we replace the source object with an object of similar geometry and reuse the trained policy without fine-tuning. The objects have similarly located origins at the bottom of their meshes, and we place these origins at the same position on the table.

B.5 Stage-wise Evaluation of DeVI

Our pipeline chains video generation, human reconstruction, object tracking, and policy learning, so errors in one stage may propagate to the next. As generated videos lack paired 3D ground truth, we report a stage-specific metric with an explicit success threshold for each stage on the same 15 generated videos. Due to resource constraints, we restrict the comparisons in Tab. 1, Tab. 5, and Tab. 8 to a shared 10-video subset of these 15 videos. The subset is fixed before running the baseline and ablation experiments, supporting fair comparisons across methods. As shown in Tab. 9, Tab. 10, Tab. 11, and Tab. 12, the stage-wise evaluation of DeVI covers all 15 videos.

Video Generation. We conduct a user study in which 71 participants rate each generated video from 1 to 5 for naturalness, physical validity, text control, and overall quality. We regard a video as successful when all four mean scores reach 4.0. Under this criterion, 13 of 15 videos succeed. The failures of Apple and Coke reflect missing object state changes, as the eaten apple shows no bite mark and the Coke bottle stays capped. These cases limit visual realism but not rigid-object HOI imitation, which is our intended use.

Human Reconstruction. We report the 2D reprojection error of the reconstructed SMPL-X joints and the average finger-to-object distance at annotated contact frames. With failure defined as a hand MPJPE above 10 px or a human-object distance above 50 mm, all 15 videos succeed. This shows

that our Visual HOI Alignment produces human motion consistent with the observed 2D joints and with the 3D object.

Object Tracking. We track each point forward and then backward and measure the displacement from its start over points with visibility above 0.1, which reports drift without ground truth. We also report $R_{v>\tau}$, the percentage of points retained at visibility thresholds $\tau = 0.1$ and $\tau = 0.75$. With failure defined as a mean forward-backward error above 10 px, 13 of 15 videos succeed.

Policy Learning. We compare the projected vertices of the simulated object with the 2D tracks, compare the simulated and reference joints, and measure penetration as the bidirectional depth from the human and object SDFs. With failure defined as an object 2D error above 20 px or a human MPJPE above 50 mm, policy learning succeeds on 14 of 15 videos. Book is the only failure, which we attribute to the difficulty of inserting fingers beneath its flat underside. Notably, Camera and Hamer fail at the tracking stage yet still yield successful policies, since the reconstructed human motion and the contact reward jointly constrain the interaction and let the policy settle on an object trajectory that best agrees with the noisy tracks. This recovery indicates that a failure in one stage does not necessarily propagate to the final policy.

Object	User Study ↑				S/F
	Nat.	Phys.	Text	Overall	
Apple	3.75	3.73	3.90	3.70	F
Book	4.17	4.46	4.48	4.28	S
Cactus	4.23	4.37	4.51	4.28	S
Camera	4.15	4.32	4.49	4.23	S
Coke	3.69	3.44	4.06	3.61	F
Cookpot	4.03	4.08	4.34	4.06	S
Garbage	4.28	4.35	4.49	4.32	S
Hamer	4.07	4.31	4.30	4.17	S
Mug	4.30	4.20	4.28	4.13	S
Pot	4.27	4.42	4.39	4.28	S
Softdrink	4.30	4.41	4.56	4.31	S
Strawhat	4.31	4.27	4.42	4.27	S
Teapot	4.39	4.51	4.41	4.37	S
Trophy	4.11	4.34	4.42	4.13	S
Wok	4.13	4.23	4.38	4.14	S
Mean	4.14	4.23	4.36	4.15	13/15

Table 9: **Video Generation.** Video generation fails when any of the four scores falls below 4.0.

Object	MPJPE (px) ↓			HOI Distance (mm) ↓	S/F
	Body	Hand	All		
Apple	15.3	3.16	7.09	19.2	S
Book	53.2	5.33	20.8	31.3	S
Cactus	27.0	2.98	10.7	15.8	S
Camera	32.4	7.04	15.2	19.2	S
Coke	29.1	3.39	11.7	10.9	S
Cookpot	29.9	3.86	12.3	16.4	S
Garbage	33.9	2.88	12.9	14.3	S
Hamer	30.9	3.65	12.4	28.0	S
Mug	43.2	4.67	17.1	19.2	S
Pot	34.6	5.37	14.8	27.0	S
Softdrink	18.6	2.38	7.59	27.2	S
Strawhat	27.2	4.38	11.8	14.1	S
Teapot	32.7	6.19	14.8	33.9	S
Trophy	32.0	4.44	13.3	45.3	S
Wok	24.1	4.78	11.0	24.4	S
Mean	31.0	4.30	12.9	23.1	15/15

Table 10: **Human Reconstruction.** Human reconstruction fails when hand MPJPE exceeds 10 px or HOI distance exceeds 50 mm.

Object	FB Err. (px) ↓	Retained (%) ↑		S/F
		$R_{v>0.1}$	$R_{v>0.75}$	
Apple	6.75	68.7	0.00	S
Book	5.71	100	71.2	S
Cactus	1.95	93.9	67.3	S
Camera	26.4	99.5	43.7	F
Coke	4.70	100	3.10	S
Cookpot	6.65	95.6	37.2	S
Garbage	3.99	89.1	45.5	S
Hamer	10.6	93.0	0.00	F
Mug	4.90	82.2	0.00	S
Pot	1.00	99.6	87.0	S
Softdrink	5.29	25.7	7.70	S
Strawhat	8.02	88.0	0.00	S
Teapot	2.01	83.8	58.6	S
Trophy	1.10	100	83.1	S
Wok	3.83	45.5	22.7	S
Mean	6.19	84.3	35.1	13/15

Table 11: **Object Tracking.** Object tracking fails when the forward-backward error exceeds 10 px.

Object	Tracking Err. ↓		Penetration (mm) ↓	S/F
	Object (px)	Human (mm)		
Apple	9.36	26.4	1.42	S
Book	356	612	0.545	F
Cactus	5.70	39.8	3.90	S
Camera	4.83	27.9	7.09	S
Coke	3.72	25.9	2.12	S
Cookpot	10.6	19.8	1.70	S
Garbage	6.63	29.5	2.88	S
Hamer	3.84	26.2	1.77	S
Mug	4.95	27.3	1.65	S
Pot	3.36	34.9	3.37	S
Softdrink	6.66	42.4	0.860	S
Strawhat	15.6	42.8	2.32	S
Teapot	3.55	27.3	0.950	S
Trophy	3.29	31.8	3.16	S
Wok	13.6	34.4	1.90	S
Mean	30.1	69.9	2.38	14/15

Table 12: **Policy Learning.** Policy learning fails when object tracking error exceeds 20 px or human tracking error exceeds 50 mm.

C Limitations and Future Work

C.1 Perspective Artifacts in Video Diffusion

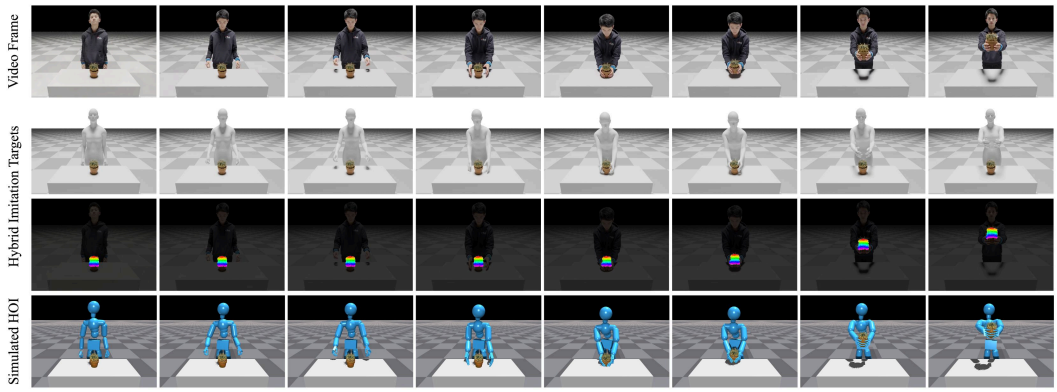
While we render a checkered grid floor alongside the scene to provide perspective cues to the video diffusion model, the model often does not produce results with perfect perspective. For example, when a human moves their hand toward the camera, the hand may appear relatively larger or smaller than it should. Such perspective artifacts introduce depth-direction errors in our Visual HOI Alignment. While our HOI loss (based on Chamfer distance) minimizes these errors near the frame where contact starts, errors in the remaining frames may reduce the naturalness of the reconstructed motion. In particular, this error becomes significant in interactions that require precise target placement (*e.g.*, putting a baseball into a small cup). As a future direction, we can consider leveraging a multi-view video diffusion model to reduce this depth error.

C.2 Visibility Loss in 2D Object Tracking

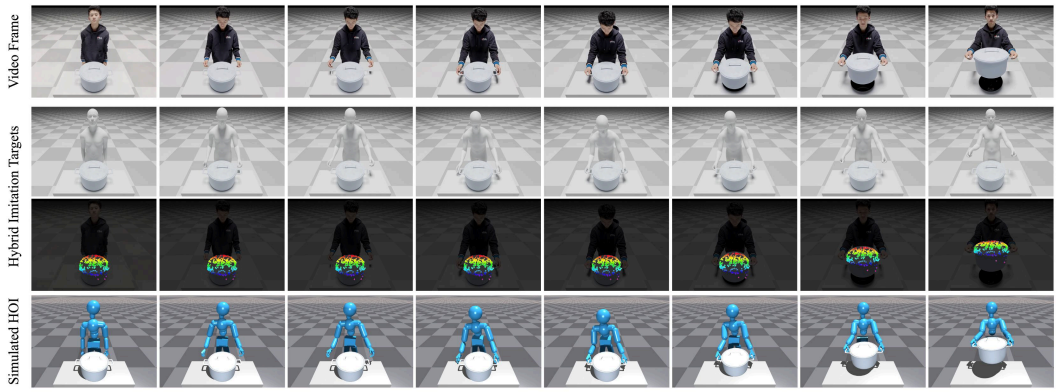
While our 2D object tracks avoid the unreliable reconstruction of 3D object motion, the tracks are defined on the object vertices visible in the first frame, so the target weakens once those vertices leave the view. This happens most often under a rapid change of object orientation, where the tracked surface rotates away from the camera within a few frames and the tracker either loses the points or drifts. As shown in Tab. 11, Camera shows the largest forward-backward error among all videos, as the person lifts the camera and turns it toward the face. Four videos retain no point at the stricter visibility threshold, leaving the object target with little information during the corresponding interval. Although we filter heavily occluded vertices using the occlusion mask of the tracker, filtering does not recover a surface that rotates out of view. The same limitation applies to in-hand reorientation and to an object that flips completely, where no vertex from the first frame stays visible. While the policy still succeeds for Camera in Tab. 12 because the human reference and the contact reward constrain the interaction, the object target contributes little during such an interval.

C.3 Unified Policy Learning

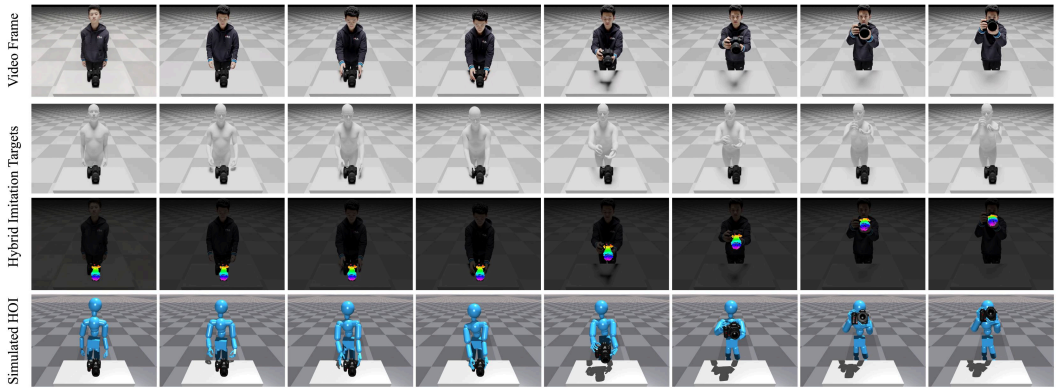
While our policy learning is fully automatic and applies the same rewards and training setup to every video without per-video tuning, DeVI still trains a separate policy for each generated video, which makes policy learning the most expensive stage of our pipeline. Learning a single policy conditioned directly on 2D object tracks remains unexplored, so our scope is limited to obtaining diverse functional manipulations rather than a controller that covers every interaction. As a future direction, conditioning a unified controller on the hybrid imitation target would allow a new video to be imitated through inference rather than training.



"Picks up a cactus with both hands"

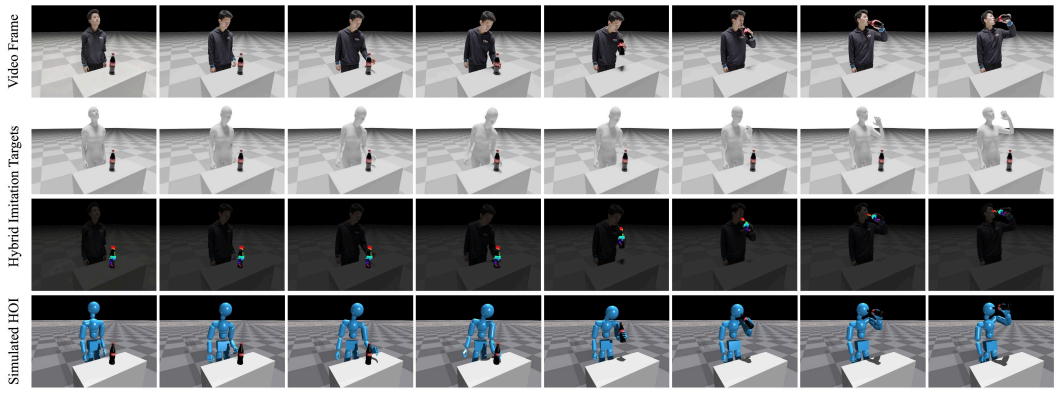


"Lifts up a pot with both hands"

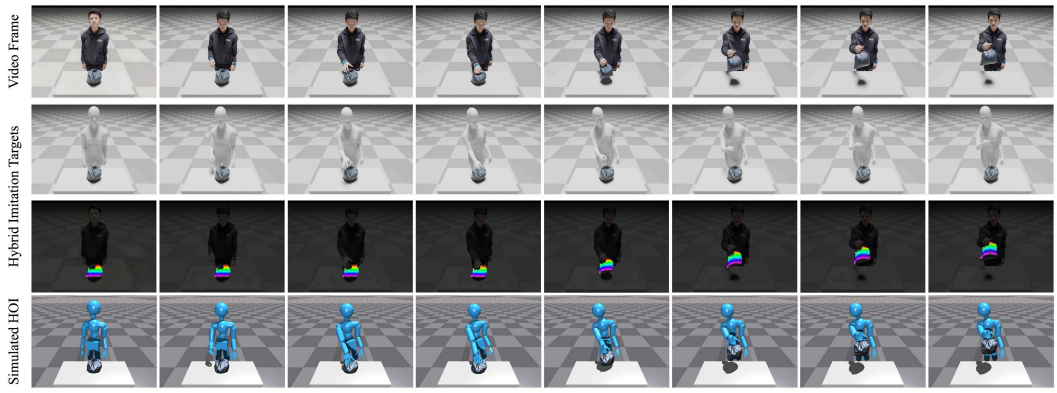


"Picks up a camera with both hands and takes a photo"

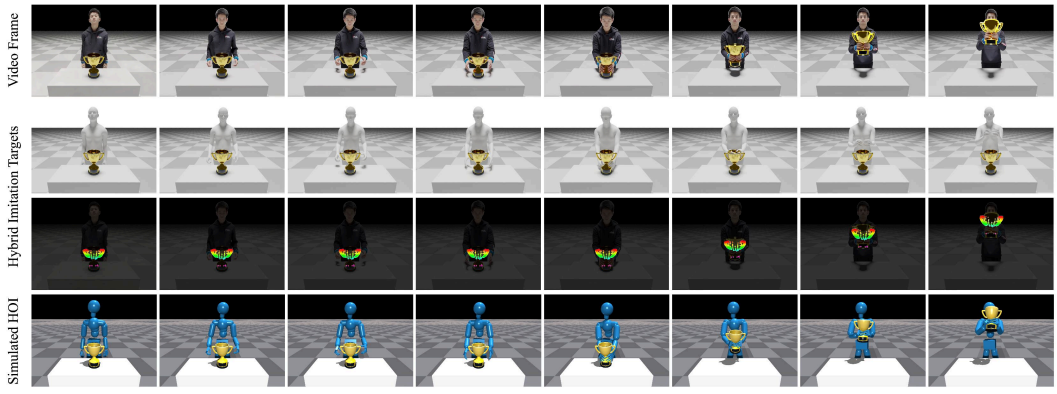
Figure 16: **Video-Motion Alignment.**



"Picks up a coke with left hand and takes a sip"



"Picks up a garbage with right hand"



"Lifts up a trophy with both hands"

Figure 17: **Video-Motion Alignment.**

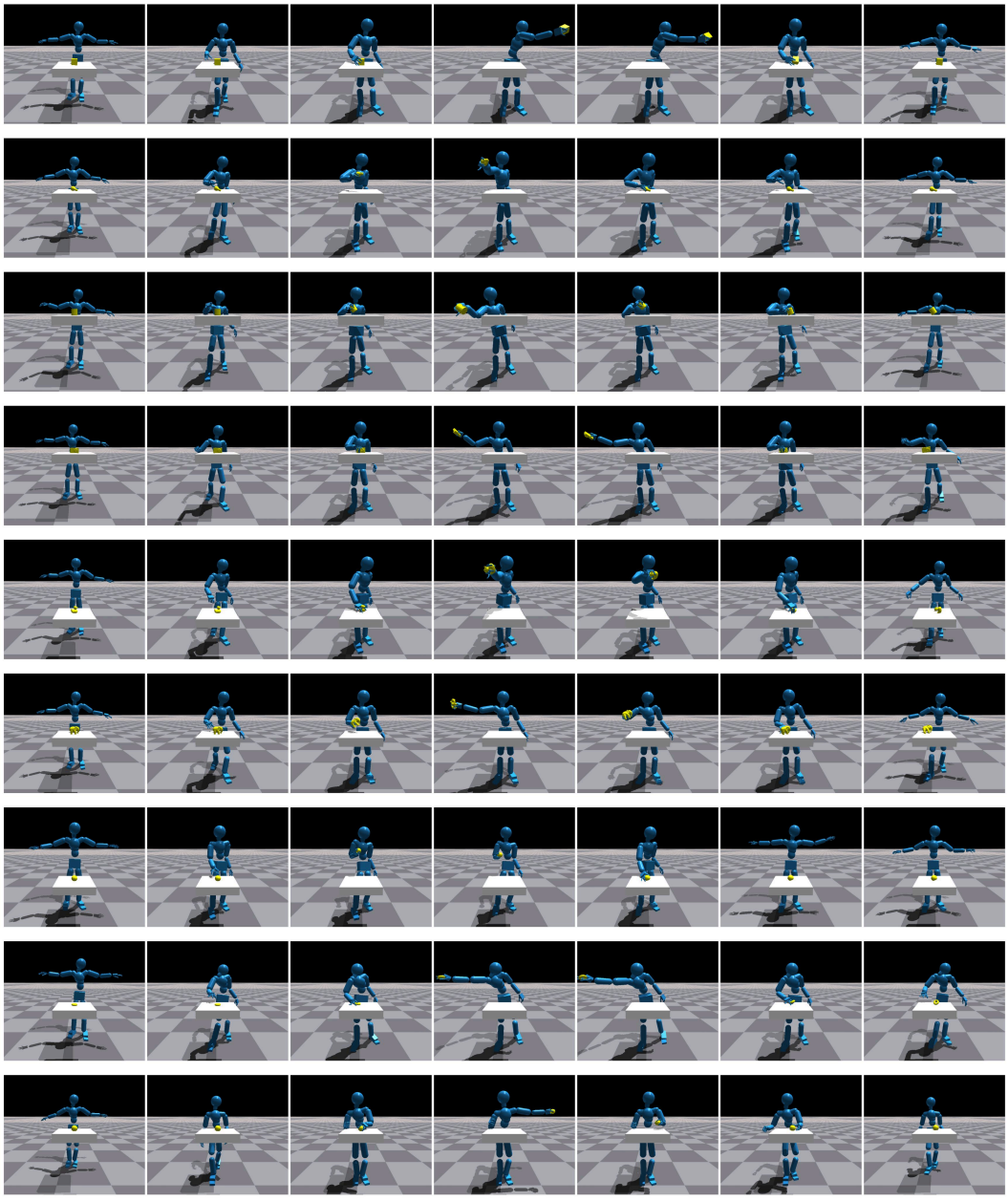


Figure 18: **Additional Results of DeVI on GRAB Dataset.**

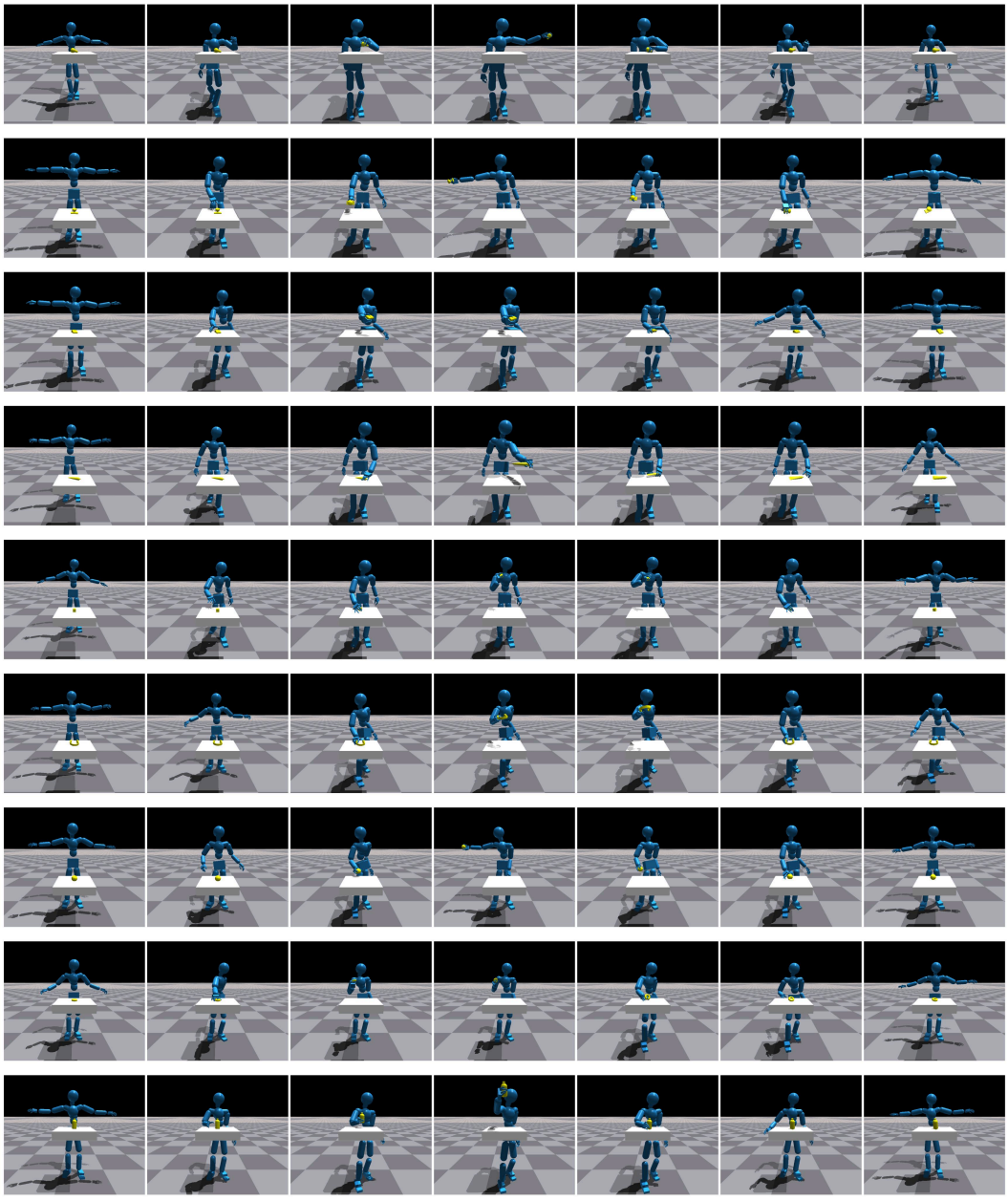


Figure 19: **Additional Results of DeVI on GRAB Dataset.**