ART2SIM: LEARNING PHYSICALLY EXECUTABLE FULL-BODY HUMANOID INTERACTION WITH ARTICULATED OBJECTS FROM VIDEO PRIORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video priors provide a scalable source of whole-body human-object interactions, yet visually coherent trajectories do not guarantee successful physical execution. This gap is especially acute for articulated objects. To open a door or pull a drawer, a humanoid must sustain contact with the correct moving link and coordinate whole-body motion to drive the object's passive joint. To bridge this gap, we introduce **Art2Sim**, a video-to-physics framework that transforms a monocular interaction video paired with an articulated asset into physically feasible full-body manipulation. Our framework consists of two stages: (1) Contact-Aware Articulated Reconstruction builds a contact-consistent kinematic reference that aligns the hand with the moving link; (2) Articulation-Guided Physics Refinement learns a residual policy on top of a pretrained whole-body tracker, using contact and joint-progress rewards to convert the reconstructed 4D HOI reference into feasible interactions. We evaluate the full pipeline on real and asset-conditioned generated videos, and the physics refiner separately on captured HOI sequences. Across the reconstruction and physics benchmarks, Art2Sim achieves state-of-the-art performance. Ablations show that our reward design is critical for sample-efficient policy learning. By grounding video priors in physical execution, our work demonstrates their potential for embodied skill learning.

1 INTRODUCTION

Articulated objects such as doors, drawers, and cabinets are ubiquitous in human environments, and manipulating them is a core capability for humanoid robots (Jiang et al., 2023; Weng et al., 2025). Learning diverse articulated interactions requires demonstrations across objects and behaviors, but motion capture and teleoperation are expensive to collect at scale. Monocular videos provide widely available human-object interaction (HOI) priors, yet existing methods mainly recover visually plausible kinematic motion without ensuring physical feasibility (Xu et al., 2021; Wang et al., 2026; Huang et al., 2026b). In parallel, physics-based HOI methods learn full-body interaction policies from motion references (Wang et al., 2023; 2025; Xu et al., 2025; 2026). Recent work further bridges monocular videos and physics-based control by reproducing reconstructed interactions in simulation (Xie et al., 2026; Huang et al., 2026a; Wu et al., 2026; Kim et al., 2026), but remains largely confined to rigid objects. Articulated objects introduce an additional challenge: joint constraints restrict object motion, requiring the humanoid to establish contact with the correct moving link and generate forces that drive the passive joint according to its kinematic constraints. Converting monocular articulated HOI videos into physically executable full-body humanoid policies remains an open challenge.

Physically executing such interactions requires three vital capabilities beyond motion tracking (Zhang et al., 2024; Weng et al., 2025; Deng et al., 2026; Zhang et al., 2026a; Zhu et al., 2026). **(1) Contact-aware reference recovery.** Reconstruction must recover the moving-link trajectory and align the manipulating hand with the link over the contact interval. **(2) Contact-mediated joint actuation.** Under contact dynamics, the humanoid must realize this hand-link relation while generating forces that advance the object's passive joint. **(3) Sustained whole-body coordination.** As the link moves, the humanoid must preserve effective contact and balance throughout the motion. As illustrated by the physics rollout in Fig. 1, opening a low oven door requires the hand to follow the

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.

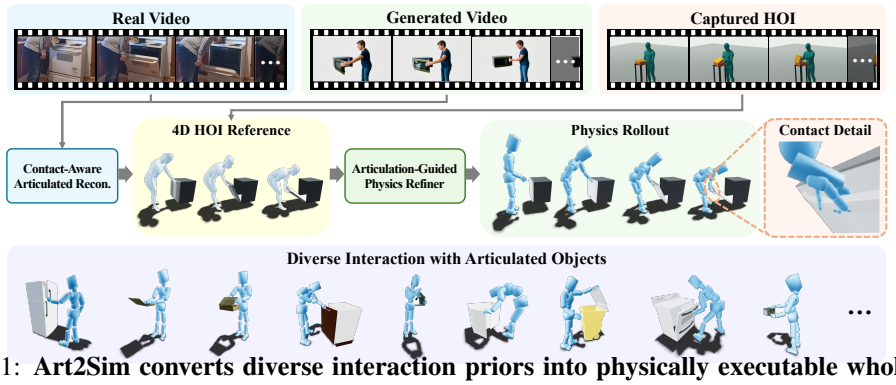

Figure 1: **Art2Sim converts diverse interaction priors into physically executable whole-body interactions with articulated objects.** Across real videos, asset-conditioned generated videos, and captured HOI references, the humanoid effectively maintains contact with the moving link and drives the passive joint in simulation. See the zoomed-in contact detail.

rotating door while the body bends to maintain reach and balance. Given the imperfect kinematic reference caused by occlusion and depth ambiguity in monocular settings, this work investigates how a full-body humanoid can robustly and effectively reproduce an articulated interaction by driving the object's passive joint through contact.

To address this problem, we introduce **Art2Sim**, a video-to-physics framework that couples contact-aware reconstruction with physics-based execution (Fig. 1). We assume each monocular video is paired with an articulated asset defining the object's geometry and kinematics. Given these inputs, **Contact-Aware Articulated Reconstruction** addresses contact-aware reference recovery by using vision-language model (VLM) contact and a grasp prior defined on the moving-link surface to refine an initial reconstruction. **Articulation-Guided Physics Refiner** ensures contact-mediated joint actuation and sustained whole-body coordination by refining this kinematic reference under contact dynamics. It builds on the pretrained PHC-X whole-body tracker with articulated fingers (Luo et al., 2024a) and articulation-aware reinforcement learning with contact and joint-progress rewards, which maintain stable hand–link contact and advance the passive joint.

We evaluate Art2Sim across articulated interactions obtained from real videos (Xu et al., 2021), asset-conditioned synthetic videos (Raistrick et al., 2024), and captured references (Fan et al., 2023). Our framework significantly outperforms existing methods in articulated HOI reconstruction and physical execution, while ablations show that the articulation-aware reward is critical for reliable contact-driven actuation. In summary, our contributions are:

- A unified video-to-physics framework for converting monocular human interaction videos into physically executable full-body humanoid interactions with articulated objects.
- A coupled reconstruction-and-control method that first recovers a contact-aware kinematic reference and then realizes it robustly under articulated-object contact dynamics.
- We conduct extensive experiments across diverse interaction sources, demonstrating consistent improvements in articulated interaction reconstruction and physical execution and validating the effectiveness of the articulation-aware design.

2 Related Work

Articulated object and interaction reconstruction. Object-centric methods recover part geometry and joint parameters from images and videos (Qian et al., 2022; Liu et al., 2025), while D3D-HOI (Xu et al., 2021) incorporates human motion cues for articulated object reconstruction. Hand-centric methods study bimanual articulated interaction reconstruction (Fan et al., 2023; Bansal et al., 2026; Wang et al., 2026), and full-body approaches typically rely on multi-view or motion-capture inputs (Jiang et al., 2023; Kim et al., 2025). Monocular world-frame reconstruction has mainly focused on rigid objects (Xue et al., 2026). These methods focus on recovering visually and kinematically plausible interactions, while the physical executability of the recovered interaction remains unexplored.

Kinematic HOI generation. Kinematic HOI generation methods synthesize human-object interactions through motion generation or optimization. Existing approaches study whole-body grasping and contact control (Taheri et al., 2022; Zhang et al., 2022), diffusion models with physics-informed corrections or contact guidance (Xu et al., 2023; Li et al., 2024), and task-driven interactions with rigid props (Hassan et al., 2023). For articulated objects, CoDA (Pi et al., 2025) and GIRAF (Zhang et al., 2026b) generate whole-body articulated manipulation kinematically, where object joint trajectories are generated or specified explicitly. These methods produce plausible articulated motions, but the object joint trajectories are not produced through humanoid-object contact dynamics.

Physics-based HOI. Physics-based HOI methods learn humanoid policies that execute interaction references under physical dynamics. Humanoid tracking has evolved from motion imitation (Peng et al., 2018; 2021) to robust tracking of noisy or partial references (Yuan et al., 2021; Luo et al., 2023; 2024b; Tessler et al., 2024). For object interaction, PhysHOI (Wang et al., 2023) introduces contact-aware imitation rewards, which are extended to broader skills and object diversity (Wang et al., 2025; Luo et al., 2024a; Xu et al., 2025). Recent works further bridge reference generation and humanoid execution through video-driven reconstruction or residual control (Xie et al., 2026; Huang et al., 2026a; Wu et al., 2026; Zhao et al., 2025). However, these methods mainly focus on rigid objects or general interaction skills and do not address contact-driven actuation of passive joints in articulated objects.

Contact-driven articulated manipulation. Contact-driven articulation has been studied with dexterous hands, robot grippers, and demonstration retargeting (Zhang et al., 2024; 2026a; Mandi et al., 2026; Kerr et al., 2024). These methods demonstrate that articulated motion can emerge from physical contact, but typically focus on specific embodiments or task settings. Recent works further investigate contact modeling and guidance: CHORD (Zhu et al., 2026) reproduces demonstrated object-level contact wrenches, while WristMimic (Yu et al., 2026) uses kinematic wrist guidance without force- or torque-level supervision. However, existing approaches do not address monocular video-driven full-body humanoid manipulation with contact-based passive joint actuation.

3 METHOD

3.1 PROBLEM FORMULATION

Given a T-frame monocular interaction video $V = \{I_t\}_{t=1}^{T}$ and a corresponding articulated asset $\mathcal{A}$, we reconstruct the observed human-object interaction and reproduce it with a full-body humanoid in a physics simulator. The asset specifies the object's canonical geometry, kinematic tree, joint axes, and limits and is shared by reconstruction and simulation. It is matched to a real video or selected before asset-conditioned generation. Generation details are provided in Sec. B.2. **Contact-Aware Articulated Reconstruction** (Sec. 3.2) constructs a contact-aware reference, and the **Articulation-Guided Physics Refiner** (Sec. 3.3) trains a reference-conditioned policy that drives the object's passive joint through contact.

The reconstruction stage maps $(V, \mathcal{A})$ to a temporally coherent interaction reference

$$\widehat{\mathcal{R}} = \mathrm{Recon}(V, \mathcal{A}) = \left(\widehat{\mathcal{R}}_{1:T}, \widehat{s}\right), \qquad \widehat{\mathcal{R}}_{1:T} = \left\{\widehat{H}_t, \widehat{T}_t^o, \widehat{q}_t\right\}_{t=1}^{T}, \tag{1}$$

where $\widehat{H}_t$, $\widehat{T}_t^o \in \mathrm{SE}(3)$, $\widehat{q}_t \in \mathbb{R}^J$, and $\widehat{s} \in \mathbb{R}_+$ denote the full-body human state, global object pose, articulated joint state, and shared object scale, respectively. Contact cues identify the manipulating hand, participating finger regions, and contact-frame set $\mathcal{C}$. They are inferred from video or derived from captured HOI geometry, and guide reconstruction alignment and the contact reward used for policy learning. The physics refiner then learns a humanoid policy π_θ conditioned on $\widehat{\mathcal{R}}_{1:T}$. At time t, the policy maps the simulator state x_t and reference window $\widehat{\mathcal{R}}_{t:t+K}$ to a humanoid action a_t,

$$a_t = \pi_\theta\left(x_t, \widehat{\mathcal{R}}_{t:t+K}\right), \qquad x_{t+1} = \mathrm{Sim}(x_t, a_t; \mathcal{A}), \tag{2}$$

where Sim advances the humanoid and articulated object under contact dynamics. The resulting rollout $\tau^{\mathrm{sim}} = \{x_t, a_t\}_{t=1}^{T}$ realizes the reconstructed interaction in simulation. We describe the reconstruction and physics stages in the following sections.

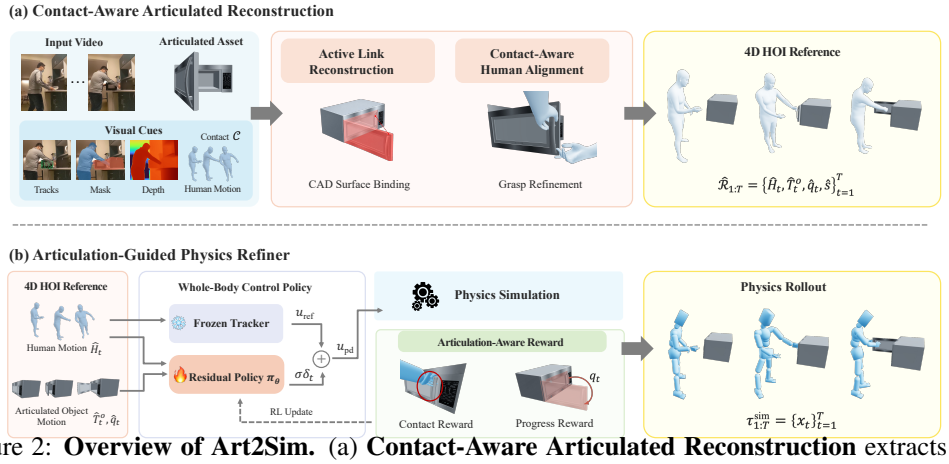

Figure 2: **Overview of Art2Sim.** (a) **Contact-Aware Articulated Reconstruction** extracts point tracks, moving-part masks, depth, human motion, and contact cues from a monocular video. It uses these cues to recover articulated motion through CAD surface binding and refine arm-hand alignment, producing a 4D HOI reference. (b) **Articulation-Guided Physics Refinement** adds a learned residual to a frozen whole-body tracker and trains the residual policy with articulation-aware rewards, producing a physics rollout that drives the object's passive joint through contact.

3.2 Contact-Aware Articulated Reconstruction

Prior articulated HOI reconstruction methods parameterize moving-part motion as per-frame $\mathrm{SE}(3)$ poses unconstrained by an explicit joint model, prioritizing geometric fidelity over physical executability (Wang et al., 2026; Huang et al., 2026b). In contrast, we jointly estimate global scale, root pose, and joint state according to the asset's kinematic tree, enforcing its prescribed joint axes and limits throughout the moving-link trajectory. We then use video-derived contact cues and a grasp prior to align the manipulating arm-hand chain with the recovered moving link.

Visual Cue Extraction. We first extract the visual cues shown in Fig. 2. SAM 2 (Ravi et al., 2024) provides the moving-part masks $\{M_t^p\}_{t=1}^T$, while CoTracker (Karaev et al., 2024) estimates 2D point trajectories and their visibility. We back-project each visible trajectory using the depth and camera intrinsics estimated by Depth Anything 3 (Lin et al., 2025), obtaining camera-frame 3D tracks $X_{t,i}$. SAM 3D Body (Yang et al., 2026) estimates the SMPL-X motion, which we temporally smooth before reconstruction. A VLM provides the manipulating hand, participating finger regions, and contact frames $\mathcal{C}$.

Active-Link Trajectory Recovery. Given the moving-part masks and lifted 3D tracks, we select the CAD link ℓ with maximum rendered overlap at an anchor frame, which identifies the corresponding joint to be actuated. We bind each anchor-frame track on ℓ to a CAD face using barycentric coordinates b_i. Forward kinematics then predicts its position at each visible frame,

$$\widehat{X}_{t,i} = \mathcal{T}_{\text{cam}}(T_t^o, q_t, \ell)\ \Phi_\ell(b_i), \qquad \mathcal{L}_{\text{part}} = \sum_{(t,i)\in\mathcal{V}} \rho\Big(\big\|X_{t,i} - \widehat{X}_{t,i}\big\|_2\Big), \tag{3}$$

Here, Φ_ℓ maps b_i to the CAD surface, $\mathcal{V}$ contains the visible track-frame pairs, and ρ is a robust penalty. Fixed surface bindings enforce a joint-consistent moving-link trajectory across frames. We initialize object root pose and scale from the silhouette and optimize root pose, scale, and joint state using $\mathcal{L}_{\text{part}}$ with silhouette, joint-limit, and temporal terms. We set $z_t = 1$ when $t \in \mathcal{C}$ and the moving-part mask changes. To suppress pre-contact joint drift, we regularize the joint state before the first active frame with

$$\mathcal{L}_{\text{static}} = \sum_{t<t_0} \|q_t - q_0\|_2^2, \qquad t_0 = \min\{t : z_t = 1\}. \tag{4}$$

Contact-Aware HOI Alignment. For each VLM-labeled contact frame $t \in \mathcal{C}$, we optimize the annotated finger regions against the transported moving-link surface. Following interaction-aware reconstruction (Jiang et al., 2023), we transport its contact surface $\mathcal{S}_\ell$ with the recovered object pose

and joint state as $\mathcal{S}_{\ell,t} = \mathcal{T}_{\text{cam}}(T_t^o, q_t, \ell)\mathcal{S}_\ell$. For the SMPL-X hand mesh $\mathcal{H}_t$ with pose θ_t^h, we define

$$\begin{aligned}\mathcal{E}_{\text{grasp}}(t) = \lambda_{\text{dis}} E_{\text{dis}}(\mathcal{H}_t, \mathcal{S}_{\ell,t}) + \lambda_{\text{fc}} E_{\text{fc}}(\mathcal{H}_t, \mathcal{S}_{\ell,t}) \\ + \lambda_{\text{pen}} E_{\text{pen}}(\mathcal{H}_t, \mathcal{S}_{\ell,t}) + \lambda_{\text{prior}} E_{\text{prior}}(\theta_t^h) + \lambda_{\text{spen}} E_{\text{spen}}(\mathcal{H}_t).\end{aligned} \tag{5}$$

The grasp prior favors proximity and force-closure-inspired contact geometry while penalizing penetration, implausible hand poses, and self-intersection. Evaluating it only on $\mathcal{S}_{\ell,t}$ ties the grasp to the moving link. $\mathcal{L}_{\text{local}}$ and $\mathcal{L}_{\text{temp}}$ further penalize spatial, velocity, and acceleration residuals of selected hand regions in the moving-link frame. Let ξ^o denote the object root pose and metric scale. We collect the motion, image, and link-conditioned alignment terms in

$$\mathcal{E}_{\text{HOI}} = \lambda_p \mathcal{L}_{\text{part}} + \lambda_m \mathcal{L}_{\text{mask}} + \lambda_c \mathcal{L}_{\text{contact}} + \sum_{t \in \mathcal{C}} \mathcal{E}_{\text{grasp}}(t) + \lambda_l \mathcal{L}_{\text{local}} + \lambda_t \mathcal{L}_{\text{temp}} + \lambda_a \mathcal{L}_{\text{anchor}}. \tag{6}$$

$\mathcal{L}_{\text{mask}}$ fits the object silhouette, $\mathcal{L}_{\text{contact}}$ attracts VLM-selected hand vertices to the moving link, and $\mathcal{L}_{\text{anchor}}$ preserves the recovered trajectory away from contact. We optimize the object and human variables in two ordered phases,

$$\xi^{o\star} = \arg\min_{\xi^o} \mathcal{E}_{\text{HOI}}(\xi^o, \bar{q}, \bar{h}), \qquad \delta h^\star = \arg\min_{\delta h_{\text{arm,hand}}} \mathcal{E}_{\text{HOI}}(\bar{\xi}^o, \bar{q}, h(\delta h)), \tag{7}$$

Here, bars denote fixed quantities, and $h(\delta h)$ applies the arm-hand residuals to the reconstructed human motion. We first optimize ξ^o with the human and articulated motion fixed, then freeze the object trajectory and refine the manipulating arm-hand chain. The second phase optimizes $\delta h_{\text{arm,hand}}$ as a temporally smooth spline over the shoulder, elbow, wrist, and hand pose of the manipulating arm, analogous to temporal inverse kinematics guided by the contact and grasp terms. Updating the chain distributes the link-relative adjustment across the limb while leaving the remainder of the reconstructed full-body motion unchanged. The resulting contact-aware kinematic reference embeds a temporally aligned, link-relative hand configuration within the reconstructed full-body motion.

3.3 Articulation-Guided Physics Refiner

The reconstruction output serves as a nominal kinematic reference for physics refinement. It specifies the target hand-link relation and articulated motion, while the policy determines the feedback-driven action adjustments required under contact dynamics. Residual control preserves whole-body tracking, while articulation-aware rewards guide the actions that sustain the intended contact and drive the passive joint.

Reference-Conditioned Residual Policy. Following residual reinforcement learning (Zhao et al., 2025), the pretrained PHC-X whole-body tracker with articulated fingers (Luo et al., 2024a) supplies the reference-tracking action, allowing the policy to focus on interaction-specific adjustments that establish contact and actuate the passive joint. The frozen tracker maps the reference to a PD target u_t^{ref}, which the residual policy augments as

$$u_t^{\text{pd}} = u_t^{\text{ref}} + \sigma\, \delta_t, \qquad \delta_t \in [-1, 1]^{153}, \tag{8}$$

where δ_t is the residual action and $\sigma = 1.0$. The policy observes

$$o_t = \left[\, x_t^h \,;\, x_t^o \,;\, q_t, \widehat{q}_{t:t+K} \,;\, a_t^{\text{teacher}} \,;\, c_t^{\text{contact}} \,\right], \tag{9}$$

where x_t^h and x_t^o are the human and object-link states, $\widehat{q}_{t:t+K}$ is the reference window of the articulated joints, a_t^{teacher} is the frozen tracker's action, and c_t^{contact} contains compressed fingertip contact forces and the poses of the finger segments and wrists relative to the object base and the moving link. These features reveal when realized contact departs from the reference hand-link relation, allowing the residual policy to adjust its action accordingly. The policy observation is specified in Sec. A.

Articulation-Aware Reward. The reward combines multiplicative tracking factors with an additive progress term,

$$r_t = r_t^{\text{body}}\, r_t^{\text{obj}}\, r_t^{\text{link}}\, r_t^{\text{contact}}\, r_t^{\text{part}}\, r_t^{\text{part-vel}} + r_t^{\text{prog}}, \tag{10}$$

Each factor scores a tracking error E as $\exp(-\lambda E)$ with a term-specific weight: r^{body} and r^{obj} track the human motion and the object root, r^{link} the moving-link pose, and r^{part} and $r^{\text{part-vel}}$ the position and velocity of the passive joint. r^{contact} scores contact localization over the 30 finger segments of both hands, rewarding the annotated segments for touching the object during the annotated contact

frames. This term transfers the fine-grained contact annotation to policy learning, while the progress term below measures whether the realized contact produces the intended articulation.

Tracking alone does not reward joint-directed force application: once the joint lags, r^{part} is near zero whether or not the contact moves it. The progress term rewards reference-directed joint displacement from the simulated state at phase entry. We segment the reference joint trajectory into opening, closing, and stationary phases. In a moving phase with direction $d \in \{-1, 1\}$, amplitude Δ, and simulated entry state q_e, the credited displacement is

$$\bar{p}_t = \text{clip}\left(\frac{d\,(q_t - q_e)}{\Delta},\ 0,\ \text{clip}\left(\frac{d\,(\widehat{q}_t - q_e)}{\Delta}, 0, 1\right)\right), \tag{11}$$

where q_t and $\widehat{q}_t$ are the simulated and reference joint states; the upper bound is the reference's own displacement from the same state. The reward credits this displacement and penalizes running ahead of the reference,

$$r_t^{\text{prog}} = \Big[w_{\text{p}}\,\bar{p}_t - w_{\text{o}}\,o_t\Big]_+, \qquad o_t = \Big[\frac{d\,(q_t - \widehat{q}_t)}{\Delta} - \epsilon\Big]_+, \tag{12}$$

where $[\cdot]_+ = \max(\cdot, 0)$ and o_t measures how far the joint leads the reference beyond a tolerance ϵ. The term is zero in stationary phases. It credits displacement that follows the demonstration and discounts displacement produced by an impact, which runs ahead of the reference.

4 Experiments

4.1 Experimental Setup

Evaluation settings. D3D-HOI (Xu et al., 2021) contains 256 real monocular videos across nine articulated-object categories, each paired with a matched PartNet-Mobility CAD model (Mo et al., 2019). We evaluate reconstruction on all 256 videos. For end-to-end video-to-physics evaluation, we randomly select five videos from each category, yielding 45 cases, and pass the references produced by our reconstruction module to physics refinement. We also generate interaction videos with Kling (KLING AI Team, 2024) and Seedance 2.5 (ByteDance Seed Team, 2025), conditioned on articulated Infinigen assets (Raistrick et al., 2023; 2024). The same assets are used for reconstruction and simulation. For physics-only evaluation, we sample 40 captured sequences from ARCTIC (Fan et al., 2023) and use them as controlled references. Dataset composition, object categories, articulated assets, and case-selection details are provided in Sec. B.

Baselines. For reconstruction, we compare with D3D-HOI (Xu et al., 2021), ArtHOI-Hand (Wang et al., 2026), and ArtHOI-GS (Huang et al., 2026b). Both ArtHOI variants are adapted to the matched-CAD setting. Implementation details are provided in Sec. A.4. For physics execution, we compare with InterMimic (Xu et al., 2025), HDMI (Weng et al., 2025), and CoDA (Pi et al., 2025). Methods within a physics setting receive the same frozen reference. This control separates an advantage in reference construction from an advantage in physically executing a given reference.

Metrics. We report object orientation, location, dimensions, and articulated joint-state errors to assess reconstruction. For generated videos, we report reference-free geometric and temporal metrics. For physics, success, human and object tracking errors, articulation progress, and contact coverage measure complementary conditions for execution. In particular, progress tests whether the passive joint moves, and contact coverage tests whether it is moved through the intended hand and link interaction. Full definitions and evaluation details are in Sec. B.

4.2 Articulated HOI Reconstruction

We evaluate reconstructed references on real D3D-HOI videos and asset-conditioned generated videos. **(1) D3D-HOI.** Table 1 compares object pose, scale, and joint-state recovery. Our full model obtains the lowest location and dimension errors, 28.6 cm and 9.2 cm, respectively, among the compared methods. Relative to Ours w/o HOI Align, the full model improves location and dimension recovery while orientation and joint-state errors show a trade-off. These metrics quantify the geometric fidelity of the reconstructed reference. Figure 3 compares Ours w/o HOI Align with Ours

Table 1: **Articulated 4D HOI reconstruction on D3D-HOI (Xu et al., 2021) benchmarks.** Given the benchmark-provided matched CAD model, we evaluate recovery of object pose, scale, and articulated joint state from each monocular video. † denotes task adaptation.

Method	Orient. (deg)↓	Loc. (cm)↓	Dim. (cm)↓	Revolute (deg)↓	Prismatic (cm)↓
D3D-HOI (Xu et al., 2021)	18.022	29.772	12.204	16.235	3.402
ArtHOI-Hand† (Wang et al., 2026)	10.928	31.942	10.454	20.457	7.177
ArtHOI-GS† (Huang et al., 2026b)	10.942	32.018	10.454	19.770	5.201
Ours w/o HOI Align	9.666	32.658	9.730	12.197	2.514
Ours	10.816	28.553	9.160	13.028	3.112

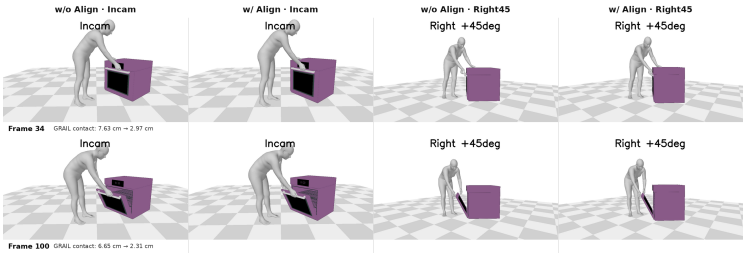

Figure 3: **Effect of HOI alignment on grasp reconstruction.** We compare Ours w/o HOI Align with Ours at representative frames, highlighting differences in the reconstructed hand-link grasp.

at representative contact frames, focusing on differences in the reconstructed hand-link grasp. **(2) Generated video priors.** Table 2 reports reference-free geometric and temporal metrics. The full model achieves the lowest contact error (6.45 cm) and the highest Part Mask IoU (0.4572), whereas Ours w/o HOI Align yields the lowest human and object smoothness errors. Generation and filtering details are provided in Sec. B.2.

4.3 Physics Execution

We evaluate the physics refiner by holding the reference fixed within each comparison and measuring whether a policy preserves full-body motion, realizes the intended hand-link contact, and drives the object's passive joint.

Reconstructed references. On D3D-HOI, we reconstruct each video once and provide the resulting frozen reference to every physics method.

Generated video priors. We execute the generated-video references from Tab. 2 with the physics refiner. Representative executions are shown in Fig. 1 and the supplementary video.

Captured references. Each physics method receives the same captured ARCTIC references (Fan et al., 2023). The benchmark contains 40 sequences from 10 object categories at 30 Hz, whose objects translate and rotate freely, so we also report object-root errors. The comparison separates body tracking from articulated execution. InterMimic tracks the body and the object root most accurately, yet its progress stays below 0.5 and no sequence succeeds: the hand follows the reference without driving the passive joint. HDMI behaves similarly. CoDA moves the joint more often but at the cost of the body, with frequent falls and a body error several times larger. Art2Sim attains the highest success, progress, and contact coverage with a body error second only to InterMimic. Body tracking alone is therefore not sufficient; success requires contact that drives the passive joint, which the ablation below attributes to the reward design.

4.4 Ablations

We ablate the physics refiner by adding its learning components cumulatively while holding the reference, asset, training budget, and evaluation protocol fixed. Rows without fine-grained contact use a hand-level contact term that only requires the reference hand to touch the object.

Table 2: **Articulated 4D HOI reconstruction on synthetic videos.** Following Xie et al. (2026), we report reference-free geometric and temporal metrics. † denotes task adaptation.

Method	Contact (cm)↓	Part Mask IoU↑	Human Smo. ($\times 10^{-2}$)↓	Object Smo. ($\times 10^{-2}$)↓
ArtHOI-GS† (Huang et al., 2026b)	13.72	0.3982	0.3567	0.1329
Ours w/o HOI Align	11.66	0.4566	0.1750	0.0706
Ours	6.45	0.4572	0.1755	0.0802

Table 3: **End-to-end video-to-physics on D3D-HOI Xu et al. (2021).** We evaluate whether the reference reconstructed from each video and its matched CAD asset can be physically executed. All methods receive the same frozen reference. † denotes task adaptation. SR, Prog., and Cov. are defined as in Tab. 5; AAE is reported per joint type because the benchmark mixes revolute and prismatic objects, and object-root errors are omitted because the object base is fixed. Each metric averages the method's available rollouts (InterMimic 35, HDMI 18, CoDA 39, Ours 42 of the 45 scheduled). HDMI covers only revolute cases.

Method	SR (%)↑	G-MPJPE	AAE		Prog.↑	Cov.↑
		(mm)↓	Rev. (°)↓	Pris. (cm)↓		
InterMimic† (Xu et al., 2025)	2.9	285.8	28.38	13.20	0.49	0.52
HDMI† (Weng et al., 2025)	5.6	595.5	28.81	–	0.58	0.04
CoDA (Pi et al., 2025)	64.1	345.6	8.84	9.71	0.83	0.53
Ours	**90.5**	**84.4**	**5.13**	**3.21**	**0.94**	**0.86**

Residual control anchors execution to the tracker. Trained with the same observations but a zero-filled teacher action, the direct policy succeeds on 5 of 40 sequences, falls in 6, and has the highest body error and the lowest contact coverage. Predicting a residual over the frozen tracker raises success to 11 of 40 and lowers body error from 199 to 172 mm. The direct policy's low object-root translation error accompanies its weak articulation: the object is disturbed less because it is engaged less. Figure 4 shows a representative box-opening rollout in which the direct policy collapses mid-sequence while every residual variant stays upright.

Progress converts contact into articulation. Fine-grained contact transfers the distance-thresholded MoCap contact annotation to policy learning and lowers body error, but does not improve success or progress. This result shows that matching the intended contact alone does not ensure joint actuation. Adding the progress term more than doubles the success rate, from 11 to 25 of 40 sequences, and raises mean progress from 0.65 to 0.83 at a small cost in body error. The persistence-based Success B criterion (Sec. B) shows the same increase. The added articulation is produced through the intended interaction: contact coverage rises with progress, and the object-root errors are the lowest of the residual rows, so the joint is driven by the grasped link rather than by

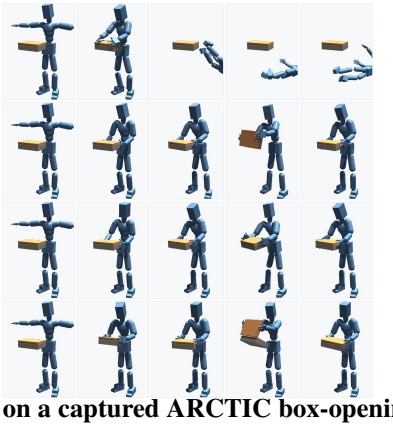

Figure 4: **Qualitative ablation on a captured ARCTIC box-opening sequence.** Rows follow the cumulative ablation of Tab. 4 from top to bottom, with five frames per rollout.

Table 4: **Cumulative ablation of the physics refiner on captured ARCTIC references.** Each row adds one component to the row above; the reference set, assets, training budget, and evaluation protocol are fixed, and SR is defined as in Tab. 5.

Method	Design choices			Overall	Human	Object			Actuation	Contact
	Residual	Fine contact	Progress	SR (%)↑	G-MPJPE (mm)↓	AAE (°)↓	Root-T (cm)↓	Root-R (°)↓	Prog. ↑	Cov. ↑
Direct policy				12.5	199.2	42.4	**16.7**	50.7	0.52	0.27
+ Residual control	✓			27.5	172.2	**30.0**	21.8	55.8	0.67	0.31
+ Fine-grained contact	✓	✓		27.5	**144.0**	37.6	26.1	55.8	0.65	0.33
+ Progress reward (Ours)	✓	✓	✓	**62.5**	149.5	36.3	20.1	**48.5**	**0.83**	**0.47**

Table 5: **Physics execution from captured ARCTIC references.** SR requires progress ≥ 0.8, mean G-MPJPE < 200 mm, mean active-frame joint error < 1 rad, and no fall; one missing HDMI rollout counts as a failure. † denotes task adaptation.

Method	SR (%)↑	G-MPJPE (mm)↓	AAE (°)↓	Root-T (cm)↓	Root-R (°)↓	Prog.↑	Cov.↑
InterMimic† (Xu et al., 2025)	0.0	**104.8**	53.3	**16.7**	51.1	0.47	0.38
HDMI† (Weng et al., 2025)	0.0	278.8	58.2	64.2	96.2	0.49	0.07
CoDA (Pi et al., 2025)	32.5	779.3	37.4	35.2	**43.8**	0.64	0.04
Ours	**62.5**	149.5	**36.3**	20.1	48.5	**0.83**	**0.47**

displacing the object. Mean AAE, which averages over all frames including holds, does not follow success. Figure 4 makes the same distinction visible: the two intermediate rows keep the hands on the box but either displace it as a whole or leave the lid closed, and only the full model opens the lid as demonstrated.

5 Conclusion

We introduce Art2Sim, a video-to-physics framework for physically executing full-body interactions with articulated objects. The central difficulty is that the object joint is passive. A video prior becomes executable only when its reconstructed hand contact remains associated with the moving link and can apply force that drives the hinge or slider in its permitted direction. Contact-Aware Articulated Reconstruction estimates this reference from the paired video and asset. Articulation-Guided Physics Refinement then adapts a whole-body tracker with residual reinforcement learning and contact and progress feedback to realize it in simulation. Experiments evaluate reconstruction fidelity and physical execution, while ablations isolate how the learning signals affect execution.

References

Siddhant Bansal, Zhifan Zhu, Shashank Tripathi, Jiahe Zhao, Michael J. Black, and Dima Damen. Towards in-the-wild egocentric 3d hand-object pose estimation. In *European Conference on Computer Vision*, 2026.

ByteDance Seed Team. Seedance: Ai video generation model, 2025. URL `https://seed.bytedance.com/en/seedance`.

Zekai Deng, Ye Shi, Kaiyang Ji, Lan Xu, Shaoli Huang, and Jingya Wang. Human-object interaction via automatically designed VLM-guided motion policy. In *International Conference on Learning Representations*, 2026. URL `https://arxiv.org/abs/2503.18349`.

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12943–12954, 2023.

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

Dingbang Huang, Etienne Vouga, Qixing Huang, and Georgios Pavlakos. Recovering physically plausible human-object interactions from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026a. URL https://arxiv.org/abs/2606.05359.

Zihao Huang, Tianqi Liu, Zhaoxi Chen, Shaocong Xu, Saining Zhang, Lixing Xiao, Zhiguo Cao, Wei Li, Hao Zhao, and Ziwei Liu. Arthoi: Articulated human-object interaction synthesis by 4d reconstruction from video priors. *arXiv preprint arXiv:2603.04338*, 2026b. URL https://arxiv.org/abs/2603.04338.

Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Zhiyuan Zhang, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker: It is better to track together. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2307.07635.

Justin Kerr, Chung Min Kim, Mingxuan Wu, Brent Yi, Qianqian Wang, Ken Goldberg, and Angjoo Kanazawa. Robot see robot do: Imitating articulated object manipulation with monocular 4d reconstruction. In *Conference on Robot Learning (CoRL)*, 2024. URL https://arxiv.org/abs/2409.18121.

Hyeonwoo Kim, Jeonghwan Kim, Kyungwon Cho, and Hanbyul Joo. DeVI: Physics-based dexterous human-object interaction via synthetic video imitation. *arXiv preprint arXiv:2604.20841*, 2026. URL https://arxiv.org/abs/2604.20841.

Jeonghwan Kim, Jisoo Kim, Jeonghyeon Na, and Hanbyul Joo. ParaHome: Parameterizing everyday home activities towards 3d generative modeling of human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

KLING AI Team. Kling image-to-video model, 2024. URL https://klingai.com/image-to-video/.

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

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025. URL https://arxiv.org/abs/2511.10647.

Yu Liu, Baoxiong Jia, Ruijie Lu, Chuyue Gan, Huayu Chen, Junfeng Ni, Song-Chun Zhu, and Siyuan Huang. VideoArtGS: Building digital twins of articulated objects from monocular video. *arXiv preprint arXiv:2509.17647*, 2025. URL https://arxiv.org/abs/2509.17647.

Zhengyi Luo, Jinkun Cao, Alexander Winkler, Kris Kitani, and Weipeng Xu. Perpetual humanoid control for real-time simulated avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Zhengyi Luo, Jinkun Cao, Sammy Christen, Alexander Winkler, Kris Kitani, and Weipeng Xu. Omnigrasp: Grasping diverse objects with simulated humanoids. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a.

Zhengyi Luo, Jinkun Cao, Josh Merel, Alexander Winkler, Jing Huang, Kris Kitani, and Weipeng Xu. Universal humanoid motion representations for physics-based control. In *International Conference on Learning Representations*, 2024b. URL https://arxiv.org/abs/2310.04582.

Zhao Mandi, Yifan Hou, Dieter Fox, Yashraj Narang, Ajay Mandlekar, and Shuran Song. DexMachina: Functional retargeting for bimanual dexterous manipulation. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2505.24853.

Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 909–918, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Mo_PartNet_A_Large-Scale_Benchmark_for_Fine-Grained_and_Hierarchical_Part-Level_3D_CVPR_2019_paper.html.

Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions on Graphics*, 37(4), 2018. URL https://arxiv.org/abs/1804.02717.

Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. AMP: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics*, 40 (4), 2021. URL https://arxiv.org/abs/2104.02180.

Huaijin Pi, Zhi Cen, Zhiyang Dou, and Taku Komura. CoDA: Coordinated diffusion noise optimization for whole-body manipulation of articulated objects. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Shengyi Qian, Linyi Jin, Chris Rockwell, Siyi Chen, and David F. Fouhey. Understanding 3d object articulation in internet videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1599–1609, 2022.

Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, Alejandro Newell, Hei Law, Ankit Goyal, Kaiyu Yang, and Jia Deng. Infinite photorealistic worlds using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12630–12641, 2023. URL https://arxiv.org/abs/2306.09310.

Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, Zeyu Ma, and Jia Deng. Infinigen indoors: Photorealistic indoor scenes using procedural generation. *arXiv preprint arXiv:2406.11824*, 2024. URL https://arxiv.org/abs/2406.11824.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloé Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL https://arxiv.org/abs/2408.00714.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. GOAL: Generating 4d whole-body motion for hand-object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Chen Tessler, Yunrong Guo, Ofir Nabati, Gal Chechik, and Xue Bin Peng. MaskedMimic: Unified physics-based character control through masked motion inpainting. *ACM Transactions on Graphics*, 43(6), 2024.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109. URL https://doi.org/10.1109/IROS.2012.6386109.

Unitree Robotics. Unitree g1 humanoid, 2024. URL https://www.unitree.com/g1/.

Yinhuai Wang, Jing Lin, Ailing Zeng, Zhengyi Luo, Jian Zhang, and Lei Zhang. PhysHOI: Physics-based imitation of dynamic human-object interaction. *arXiv preprint arXiv:2312.04393*, 2023. URL https://arxiv.org/abs/2312.04393.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Zikai Wang, Zhilu Zhang, Yiqing Wang, Hui Li, and Wangmeng Zuo. Arthoi: Taming foundation models for monocular 4d reconstruction of hand-articulated-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2603.25791.

Haoyang Weng, Yitang Li, Nikhil Sobanbabu, Zihan Wang, Zhengyi Luo, Tairan He, Deva Ramanan, and Guanya Shi. HDMI: Learning interactive humanoid whole-body control from human videos. *arXiv preprint arXiv:2509.16757*, 2025. URL https://arxiv.org/abs/2509.16757.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Shengming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025. URL https://arxiv.org/abs/2508.02324.

Tianshu Wu, Xiangqi Kong, Yue Chen, Qize Yu, Hang Ye, Jia Li, Yizhou Wang, and Hao Dong. Sugar: A scalable human-video-driven generalizable humanoid loco-manipulation learning framework. *arXiv preprint arXiv:2605.20373*, 2026. URL https://arxiv.org/abs/2605.20373.

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. *arXiv preprint arXiv:2606.05160*, 2026. URL https://arxiv.org/abs/2606.05160.

Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. InterDiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Sirui Xu, Samuel Schulter, Morteza Ziyadi, Xialin He, Xiaohan Fei, Yu-Xiong Wang, and Liangyan Gui. InterPrior: Scaling generative control for physics-based human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2602.06035.

Xiang Xu, Hanbyul Joo, Greg Mori, and Manolis Savva. D3D-HOI: Dynamic 3d human-object interactions from videos. *arXiv preprint arXiv:2108.08420*, 2021. URL https://arxiv.org/abs/2108.08420.

Lixin Xue, Chengwei Zheng, Georgios Paschalidis, Chen Guo, Manuel Kaufmann, Juan Zarate, and Dimitrios Tzionas. RHINO: Reconstructing human interactions with novel objects from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2605.17014.

Xitong Yang, Devansh Kukreja, Don Pinkus, Anushka Sagar, Taosha Fan, Jinhyung Park, Soyong Shin, Jinkun Cao, Jiawei Liu, Nicolas Ugrinovic, Matt Feiszli, Jitendra Malik, Piotr Dollár, and Kris Kitani. SAM 3d body: Robust full-body human mesh recovery. *arXiv preprint arXiv:2602.15989*, 2026. URL https://arxiv.org/abs/2602.15989.

Wongyun Yu, Youngwoon Kim, and Minsu Cho. WristMimic: Full-body humanoid control with wrist-guided manipulation. In *European Conference on Computer Vision*, 2026.

Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. SimPoE: Simulated character control for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Hui Zhang, Sammy Christen, Zicong Fan, Luocheng Zheng, Jemin Hwangbo, Jie Song, and Otmar Hilliges. ArtiGrasp: Physically plausible synthesis of bi-manual dexterous grasping and articulation. In *International Conference on 3D Vision (3DV)*, 2024.

Tianshan Zhang, Yijia Duan, Yanjun Li, Zeyu Zhang, and Hao Tang. DragMesh-2: Physically plausible dexterous hand-object interaction with articulated objects. *arXiv preprint arXiv:2606.15133*, 2026a. URL https://arxiv.org/abs/2606.15133.

Xiaohan Zhang, Bharat Lal Bhatnagar, Vladimir Guzov, Sebastian Starke, and Gerard Pons-Moll. COUCH: Towards controllable human-chair interactions. In *European Conference on Computer Vision*, 2022.

Xiaohan Zhang, Sebastian Starke, Alexander Winkler, Federica Bogo, Samir Aroudj, and Yuting Ye. GIRAF: Towards generalizable human interactions with articulated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2026b. URL https://arxiv.org/abs/2607.07880.

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. *arXiv preprint arXiv:2510.05070*, 2025. URL https://arxiv.org/abs/2510.05070.

Xinghao Zhu, Zixi Liu, Shalin Jain, Chenran Li, Milad Noori, Michael Andres Lin, Huihua Zhao, John Welsh, Mrinal Verghese, Wei Liu, Tingwu Wang, Xingye Da, Zhengyi Luo, Vishal Kulkarni, Naema Bhatti, Yuke Zhu, Linxi Fan, Bowen Wen, Danfei Xu, Soha Pouya, and Yan Chang. Learning dexterous manipulation using contact wrench guidance from human demonstration. *arXiv preprint arXiv:2607.00033*, 2026. URL https://arxiv.org/abs/2607.00033.

This supplementary material provides additional implementation details, experimental settings, and results. Sec. A describes the implementation of our reconstruction and physics-based execution components, as well as the baselines. Sec. B provides the datasets, evaluation protocol, metrics, and statistical analysis. Sec. C presents additional quantitative and qualitative results. Finally, Sec. D discusses limitations, future work, and broader impact.

A Implementation Details

A.1 Reconstruction

Given a video-asset pair $(V, \mathcal{A})$, reconstruction estimates $\widehat{\mathcal{R}}_{1:T}$, with $\mathcal{A}$ defining the canonical geometry and kinematics. We parameterize the object state by a world 6D pose $\widehat{T}_t^o$, a vector of revolute (hinge) and prismatic (slider) joint values $\widehat{q}_t$, and a scalar scale $\widehat{s}$.

Preprocessing pipeline. Before reconstruction, we extract the masks, point tracks, depth, human motion, and contact cues shown in Fig. 2.

Masks and point tracks. SAM 2 (Ravi et al., 2024) propagates an anchor-frame prompt to obtain the moving-part masks $\{M_t^p\}_{t=1}^T$. CoTracker (Karaev et al., 2024) predicts 2D trajectories $u_{t,i}$ and visibility for points sampled on the moving part.

Depth lifting. Depth Anything 3 (Lin et al., 2025) estimates per-frame depth D_t and camera intrinsics K_t. For each visible observation, we back-project the tracked pixel into the camera frame as $X_{t,i} = D_t(u_{t,i})K_t^{-1}\bar{u}_{t,i}$, where $\bar{u}_{t,i}$ is the homogeneous image coordinate. These lifted trajectories provide the 3D track observations used in Eq. (3).

Human motion and contact. SAM 3D Body (Yang et al., 2026) estimates a camera-frame SMPL-X sequence. We temporally smooth the sequence to suppress framewise jitter before aligning it with the object and world frame. A vision–language model identifies the manipulating hand, participating finger regions, and contact frames. We transfer these contact annotations to the corresponding vertices of the selected CAD link.

Asset initialization and world frame. We initialize the asset at an anchor frame by rendering candidate root poses and scales and comparing their silhouettes with the observed object mask. The support plane of the initialized asset defines the world up direction, after which the human and object are expressed in a common frame.

We optimize object root rotation and translation, joint values, and scale with the track, silhouette, joint-limit, contact, and temporal terms. After global optimization, we refine the tracks and joint trajectory and temporally smooth the recovered joint states before exporting the reference.

Human-object alignment. The alignment proceeds in two stages. In the object stage we optimize the object 6D pose, scale, and articulation under the visual and contact terms while the human stays fixed. In the human-only stage we refine the active arm chain, the wrist, elbow, shoulder, and hand pose, while the object stays fixed. The full objective is the weighted sum in Eq. (6). It combines active-part track consistency (Eq. (3)), silhouette evidence, contact and grasp geometry, and temporal anchors. The no-contact prior in Eq. (4) prevents pre-interaction joint drift. The variant without HOI alignment skips the human-only stage and keeps the object-only alignment.

A.2 Physics-Based Execution

The simulator and material settings are summarized in Tab. 6.

The refiner receives the reference $\widehat{\mathcal{R}}_{1:T}$ and the articulated asset through a common interface. The simulator state x_t contains the humanoid state, object-root and link poses, and passive joint values. D3D-HOI fixes the object base and disables object gravity; ARCTIC keeps a dynamic object root and gravity. Neither benchmark actuates the object joints or writes their reference states back into the simulator. Initial states follow the frozen dataset-specific case definitions.

The complete observation decomposition and dimensionality are given in Tab. 7.

Table 6: **Physics simulation settings.** Values are taken from the Ours physics configurations. D3D-HOI overrides object gravity and friction relative to the shared ARCTIC configuration; articulated object joints are not actuated.

Hyperparameter	Value
Simulation step Δt	$1/60\,\mathrm{s}$
Control decimation	2 ($\Delta t_{\mathrm{ctrl}} = 1/30\,\mathrm{s}$)
Parallel environments	2048
PhysX substeps	2
Solver type	1
Position / velocity iterations	4 / 1
Contact / rest offset	0.002 / 0.0
Maximum depenetration velocity	10.0
Object density	1000.0
Object gravity	disabled (D3D-HOI); enabled (ARCTIC)
Object friction / restitution	1.0 / 0.0 (D3D-HOI); 0.9 / 0.0 (ARCTIC)
Ground static / dynamic friction	1.0 / 1.0 (D3D-HOI); 0.9 / 0.9 (ARCTIC)
Object joint drive	none (stiffness = 0, damping = 0)

Table 7: **Policy observations.** The residual policy input is formed by concatenating the live humanoid and object state, reference targets, teacher signal, contact features, and hand–object geometry listed above; these dimensions sum to 1992. The teacher-action block is zero-filled only for the direct-policy ablation.

Term	**Description**	**Dim.**	**Source**
Live state			
Body state	humanoid body observation	778	simulator
Object links	link pose and velocity features	30	simulator
Current DOF	humanoid joint positions	153	simulator
DOF velocity	humanoid joint velocities	153	simulator
Active joint	current articulated-object joint value	1	simulator
Teacher action	frozen tracker action	153	teacher
Reference and object targets			
Reference DOF	reference humanoid joint positions	153	reference
Reference active joint	reference articulated joint value	1	reference
Reference object root	root position, rotation, velocity, angular velocity	15	reference
Reference active link	active-link position, rotation, velocity, angular velocity	15	reference
Link target error	current-to-reference link pose/velocity error	30	derived
Contact and hand–object geometry			
Fingertip force	ten local compressed 3D fingertip forces	30	contact
Hand–object pose	30 finger segments and two wrists, relative to base/active links	480	derived
Total	residual-policy observation	**1992**	

The policy observes a 1992-D vector assembled from the body state, the object link state, the current, velocity, and reference humanoid DOF, the teacher action, the active joint, compressed fingertip forces, and finger and wrist poses relative to the object's base and active links. The action is the 153-D residual in Eq. (8) with $\sigma = 1.0$. The reward follows Eq. (10) and Eq. (11). The multiplicative body, object, active-link, contact, part, and part-velocity terms are exponentials of their costs. The additive progress term rewards reference-paced, contact-weighted articulation progress; D3D-HOI also penalizes excessive progress ahead of the reference.

A.3 Training Details

We train all reported policies with PPO (Schulman et al., 2017). Each case replicates the reference into 2048 parallel environments with an episode length of 300 and 16384 samples per minibatch. The residual policy is a 1024-1024-512 MLP with separate actor and critic heads and a fixed exploration variance. Evaluation reports the selected checkpoint from this fixed schedule. Rollouts use

Table 8: **PPO hyperparameters.** Values are copied from the Ours configuration. The final row records the dataset-specific schedules rather than treating the two schedules as identical.

Hyperparameter	Value
Action distribution	153-D continuous residual
Residual scale and clip	$\sigma = 1.0$, $[-1, 1]$
Policy network	separate actor/critic MLP, 1024-1024-512, ReLU
Parallel environments	2048
Episode length	300 steps
PPO horizon length H	32
Minibatch size	16384
Mini-epochs	6
Discount factor γ	0.99
GAE parameter τ	0.95
Learning rate	2×10^{-5} (constant)
PPO clip coefficient	0.2
Critic coefficient	5.0
Bounds-loss coefficient	10.0
Entropy coefficient	0.0
Exploration log standard deviation	-2.9 (fixed)
Initialization	Hybrid, probability 0.1, max fraction 0.5
Training epochs / checkpoint interval	1000 / 100

Table 9: **Comparison of reconstruction and physics methods in their native settings.** (a) We compare reconstruction methods by video input, human scope, and whether the object model is recovered or provided. (b) We compare physics methods by native task, object type, and embodiment.

(a) Articulated HOI Reconstruction

Method	Video Input	Human Scope	Object Representation
ArtHOI (Wang et al., 2026)	Monocular RGB	Hands (MANO)	Recovered articulated mesh
ArtHOI-GS (Huang et al., 2026b)	Generated video prior	Full body (SMPL-X)	Recovered 3D Gaussians
Ours	Monocular RGB / video prior	Full body (SMPL-X)	Provided articulated CAD asset

(b) Physics Execution

Method	Native Task	Object Type	Embodiment
InterMimic (Xu et al., 2025)	Whole-body HOI imitation	Rigid	SMPL-X
HDMI (Weng et al., 2025)	Whole-body control from human videos	Articulated	Unitree G1
CoDA (Pi et al., 2025)	Whole-body articulated manipulation	Articulated	SMPL-X
Ours	Video-to-physics execution	Articulated	SMPL-X

hybrid reference-state initialization during training; Tab. 8 summarizes these configuration values. Exact dependency and simulator versions, hardware, and per-run wall-clock time are reported in the released code.

A.4 Baseline Implementations

Reconstruction baselines. D3D-HOI (Xu et al., 2021) is evaluated with its author-released optimized results under the native GT-CAD protocol, using the same evaluator as our two variants. For ArtHOI† (Wang et al., 2026), empty active-part observations fall back to its IoU term without Sinkhorn.

Physics baselines. Methods marked with † in Tabs. 3 and 5 are adapted to the shared SMPL-X articulated-object benchmark. We retain their native learning objectives and report the task-facing changes needed for the matched comparison. We use the official CoDA articulated-object controller, derived from PhysHOI (Wang et al., 2023), and adapt only its interface to the shared reference, asset, and evaluation API.

Table 10: **Dataset case composition.** Numbers in parentheses are the trajectories used in the corresponding experiments, as specified by the D3D-HOI45, Infinigen summary, and ARCTIC40 manifests.

Dataset	Articulated object types (number of trajectories)
D3D-HOI / PartNet	dishwasher (5), laptop (5), microwave (5), oven (5), refrigerator (5), storage_prismatic (5), storage_revolute (5), trashcan (5), washingmachine (5)
Infinigen	cabinet (5), dishwasher (3), door (6), drawer (6), microwave (5), oven (3), refrigerator (2), window (5)
ARCTIC	box (4), capsulemachine (4), espressomachine (4), ketchup (4), laptop (4), microwave (4), mixer (4), notebook (4), phone (4), waffleiron (4)
Total	45 + 35 + 40 = 120 trajectories

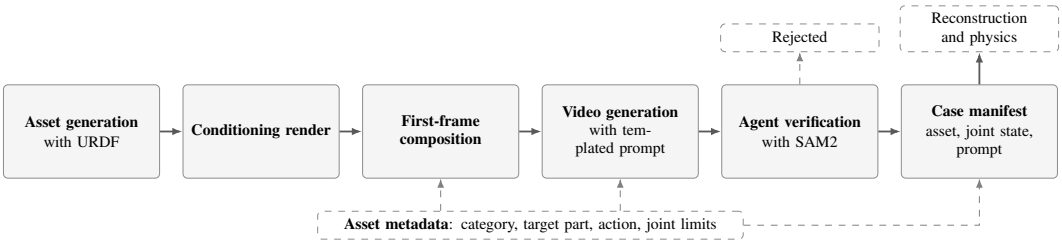

Figure 5: **Asset-conditioned video generation.** Solid arrows denote per-case data flow. Dashed arrows denote reused asset metadata and rejected videos.

B Experimental Details

B.1 Datasets

The object categories and trajectory counts used in each benchmark are listed in Tab. 10. D3D-HOI pairs each real RGB sequence with a matched PartNet-Mobility CAD model and supports reconstruction and end-to-end evaluation. ARCTIC annotations provide captured references for physics-only evaluation.

B.2 Asset-Conditioned Generated Videos

For generated-video cases, the same articulated asset conditions video generation and defines the reconstruction and simulation model. The pipeline in Fig. 5 renders a conditioning frame, generates V, and stores $(V, \mathcal{A})$ with the target part, initial joint state, and joint limits.

Assets and conditioning render. We procedurally generate articulated assets with Infinigen (Raistrick et al., 2023; 2024), sampling multiple instances per category from random seeds and exporting each instance as a URDF with visual and collision geometry. For conditioning, we render each asset in MuJoCo (Todorov et al., 2012) with every joint set to the midpoint of its limits, on a white background at 1024×1024 pixels, with the camera placed at a front or oblique preset fitted to the asset bounding box. The configuration allows the same asset to condition both opening and closing status and provides a known initial joint state for the conditioning image. The rendered object is then cropped and passed to video generation models (such as Kling and Seedance) in the reference image mode.

First-frame composition and prompting. An image-editing model (Qwen-Image-Edit (Wu et al., 2025)) composites a person and the object crop into a white-background first frame. The generation prompt is instantiated from the template in Fig. 6 by filling its per-case fields (the object category, target part, acting hand, and part motion) from the asset metadata and the action label. The template fixes a single continuous distant shot and a three-phase timeline in which the hand approaches the target part with a visible gap, first contact occurs at about 2.5 s, and the part moves only after contact.

Video Generation prompt.
Create one continuous 5-second photorealistic video from the exact supplied first frame. Keep the same single adult, the same object, their left-right positions, and the fixed distant 16:9 white-studio camera. The person's entire head, both hands, both legs, and both shoes, plus the complete object, must stay visible through every frame. Do not enlarge either subject or reframe.

0–1 s: both hands stay visibly separate from the object. The person approaches {*part*} of the {*object*}; the object and its movable part stay still.

1–2.5 s: extend the {*hand*} toward {*part*}, showing several frames with clear air between the fingers and the target. Around 2.5 s, make first contact and establish a visible grip. Only after contact, {*part motion*}, keeping the acting hand on the same part.

4–5 s: finish and hold the {*object*} in its final state. No cut, zoom, pan, camera tilt, extra person, extra object, or changed clothing.

Figure 6: **Prompt templates for asset-conditioned video generation.** Italic braces denote per-case fields filled from the asset category, the target part, and the action.

Video generation. The pipeline submits the first frame and the prompt to the video-generation interface, polls the task until completion, and downloads the resulting 5-second clip. Neither interface exposes a sampling seed. Each case receives one generation request without retries or sample selection, and its task identifier is recorded in the case manifest.

Agent-based verification. An automated verifier screens each video before reconstruction. It segments and tracks the object with SAM2 (Ravi et al., 2024) over the clip and extracts keyframes at 0.5 s intervals, giving 11 frames over 0–5 s. GPT-6 then judges the keyframes together with the object masks against four criteria: (i) the whole body and the complete object remain in view; (ii) the clip contains no shot cut or zoom; (iii) separation, first contact, and part motion occur in this order; and (iv) the moving part is the target part, its motion direction matches the action, and no part moves without contact. Only videos satisfying all four criteria are retained. Rejected cases are not regenerated, and the criteria are fixed before reconstruction and physics evaluation. Each retained video is written to a case manifest together with its asset, initial joint state, and prompt, from which the reconstruction stage reads its inputs directly.

B.3 Articulated Object Models

Each articulated asset provides the kinematic tree, the joint types, the joint limits, and the visual and collision geometry. Visual meshes drive reconstruction and rendering, while the collision geometry, mass, inertia, and joint limits drive simulation. Both share the same link names, joint axes, and root frame. Joint values remain absolute, in radians for revolute joints and meters for prismatic joints.

B.4 Evaluation Protocol

For Physics, each pre-registered case-seed pair occupies one denominator slot. A crash, NaN, timeout, missing output, incomplete reference horizon, or fall is a binary failure. Continuous errors are computed over the valid prefix and reported with valid-duration coverage and rollout availability. The ARCTIC evaluation includes 40/40 CoDA, 40/40 InterMimic, 40/40 Ours Setting 15, and 39/40 HDMI rollouts. The unavailable HDMI case is `s04/waffleiron_use_02`.

D3D-HOI video-to-physics reporting. Table 3 evaluates the shared reconstructed reference on the 45 scheduled case-seed pairs and reports the same metrics as the ARCTIC tables: SR is Success A, Prog. is the phase-weighted normalized joint displacement, and Cov. is intended-hand contact coverage, all computed by the same metric engine and thresholds. Every metric averages the rollouts available for each method (InterMimic 35, HDMI 18, CoDA 39, Ours 42 of 45; Ours covers 39 revolute and 3 prismatic cases), so rates are proportional estimates over each method's completed cases rather than schedule totals with missing entries counted as failures. The persistence-based Success B criterion yields the same 38/42 Ours successes. Object-root tracking is omitted because the object base is fixed. These execution errors are measured against the reconstructed reference; reconstruction accuracy is evaluated separately.

B.5 Evaluation and Statistical Analysis

Evaluation metrics. The common evaluator segments each reference joint trajectory into movement phases k with start/end frames (s_k, e_k) and joint j_k. We define reference-normalized articulation progress as

$$p_k = \mathrm{clip}_{[0,1]}\left[\max_{t\in\mathcal{W}_k} \frac{q^{\mathrm{sim}}_{t,j_k} - q^{\mathrm{ref}}_{s_k,j_k}}{q^{\mathrm{ref}}_{e_k,j_k} - q^{\mathrm{ref}}_{s_k,j_k}}\right], \qquad P = \frac{\sum_k L_k p_k}{\sum_k L_k}, \quad L_k = e_k - s_k + 1, \tag{13}$$

where $\mathcal{W}_k$ is the five-frame window centered on the available end of phase k. An unstarted phase contributes zero; a reference with no scoreable phase receives $P = 0$. The score uses the maximum progress within the endpoint window. Task completion requires $P \geq 0.8$, the full reference horizon, and no fall (pelvis height below $0.2\,\mathrm{m}$). Success A additionally requires mean global 22-joint H-MPJPE $< 0.2\,\mathrm{m}$ and active-frame joint MAE $< 1.0\,\mathrm{rad}$ for revolute or $< 0.10\,\mathrm{m}$ for prismatic joints. Success B instead rejects ten consecutive frames with H-MPJPE $> 0.5\,\mathrm{m}$ or joint error $> 1.5\,\mathrm{rad}/0.10\,\mathrm{m}$. Contact coverage is reported separately and does not enter Success A or B.

Human errors are measured over the valid prefix, with global and pelvis-translation-aligned errors reported separately. The joint-MAE term in Success A uses reference-active frames. The ARCTIC table instead reports all-frame AAE in degrees to measure the entire joint trajectory, including holds. D3D-HOI's all-frame joint errors are reported in revolute degrees and prismatic centimeters. Success thresholds use radians for revolute joints and meters for prismatic joints. For ARCTIC, object-root translation error is the mean world-frame position distance and root rotation error is the mean geodesic angle of the simulated and captured orientations. Root errors are reported for dynamic-root ARCTIC and omitted for fixed-base D3D-HOI. Contact coverage is an intended-contact proxy. Collider-attributed contact and penetration are not evaluated.

B.6 Metric Definitions and Sensitivity

Coordinate-frame correction. The historical Setting 15 exporter adds the saved `tracker_local_xy_origin` to simulated world positions. The root evaluator instead reads the unconverted local reference from `studio_input/reference.npz`. This mismatch creates a mean initial translation error of 0.437 m before interaction begins. Adding the same saved XY origin to every reference-root frame changes the mean Root-T error from 0.525 m to 0.222 m; the mean initial error becomes below 10^{-7} m. We apply the saved XY offset uniformly to every reference-root frame. Policy outputs, articulation, human errors, root rotation, and Success A/B remain unchanged. The corrected root reference matches InterMimic's reference within 10^{-6} m throughout all 40 clips. HDMI is evaluated with its native reference conversion. Its initial root-orientation offset is retained in the reported root errors.

Stationary-joint counterfactual. The reference-normalized score P anchors each phase at its reference start value. In a closing phase, a joint that remained closed throughout can consequently receive full phase credit. Replacing all simulated joint values by the first reference value, with all other fields fixed, gives mean $P = 0.458$ on the 40 Setting 15 references. Thus, a threshold of 0.5 can admit stationary behavior, while endpoint maximization and clipping can obscure transient overshoot.

Actual phase displacement. As a separate, threshold-free actuation diagnostic we compute

$$P_\Delta = \frac{1}{K}\sum_{k=1}^{K} \mathrm{clip}_{[0,1]}\left(\frac{q^{\mathrm{sim}}_{e_k,j_k} - q^{\mathrm{sim}}_{s_k,j_k}}{q^{\mathrm{ref}}_{e_k,j_k} - q^{\mathrm{ref}}_{s_k,j_k}}\right). \tag{14}$$

Phases whose endpoint is unavailable contribute zero. We average phases within case, then cases, assigning zero to a missing rollout. This score uses the same reference-derived phases for a given reference and gives zero for a stationary joint. We report P_Δ with AAE and target-attainment diagnostics because it remains time-aligned and clips overshoot.

Sequence-success decomposition. Table 11 shows that progress is the dominant failed condition. InterMimic and HDMI produce 39 and 31 non-falling full-horizon rollouts, respectively, yet neither

Table 11: **Sensitivity analysis of the ARCTIC evaluation metrics.** Counts use 40 scheduled cases. The P columns require full-horizon, non-falling rollouts. TC and PCR are sensitivity diagnostics.

Method	Available	No fall	A	B	$P \geq .6$	$P \geq .7$	$P \geq .8$	TC (%)	PCR (%)	All
CoDA (Pi et al., 2025)	40	18	13	11	18	17	13	44.6	22.1	5
InterMimic† (Xu et al., 2025)	40	39	0	0	4	0	0	61.6	0.6	0
HDMI†‡ (Weng et al., 2025)	39	31	0	0	3	0	0	39.3	1.3	0
Ours (Setting 15)	40	37	25	19	30	29	26	75.2	25.4	3

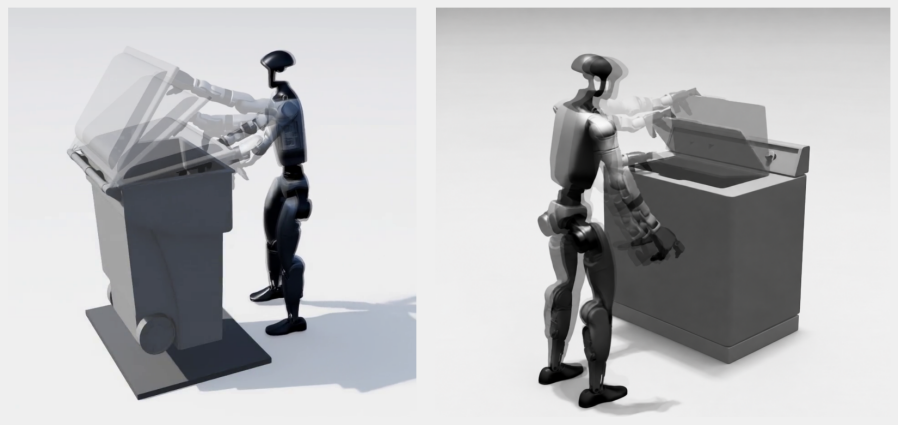

Figure 7: **Qualitative demonstration on the Unitree G1 embodiment.** Reconstructed articulated interactions are transferred to a Unitree G1, showing the applicability of our method to a different humanoid embodiment.

reaches $P \geq 0.7$. On `s01/notebook_use_02`, InterMimic articulates later than the reference, whereas HDMI attains $P \approx 0.627$ with little articulation.

Tracking-frame coverage (TC) is the fraction of scheduled frames with body error $< 0.2\,\text{m}$, joint error $< 0.5\,\text{rad}$, and pelvis height $\geq 0.2\,\text{m}$. Phase completion rate (PCR) requires each phase endpoint to lie within 20% of the reference displacement and achieve at least 80% of its directed motion. "All" requires every phase and no fall. Setting 15 passes "All" on 3/40 cases, compared with 25/40 under Success A.

Statistical analysis. Unless stated otherwise, tables report single-seed means across cases. Multi-seed analyses aggregate within case and use a paired bootstrap clustered by CAD model or object. Sensitivity analyses are descriptive.

C Additional Results

C.1 Ablation and Additional Analyses

Ablation studies. Table 4 evaluates the cumulative physics design by adding tracker-conditioned residual control, contact reward, and progress reward in sequence. All rows share the reference, asset, training budget, and evaluation protocol, and report final-budget execution metrics.

C.2 Qualitative Results

Humanoid Embodiment Demonstration. We additionally examine whether the core components of our framework can be applied to a humanoid robot. We retarget trajectories reconstructed by our method to a Unitree G1 (Unitree Robotics, 2024) and optimize their execution using the Sugar (Wu et al., 2026) codebase, with reward terms adapted from our physics-based execution stage. As shown in Fig. 7 The robot successfully executes the retargeted interactions, showing that both the reconstructed references and our reward design can be used with a humanoid embodiment. We

include this experiment as a qualitative demonstration on real hardware, while leaving a broader evaluation for future work.

D Limitations and Future Work

D.1 Limitations

Remaining failure modes. Art2Sim still fails on cases with severe occlusion or very thin contact gaps. When the hand or the manipulated part is heavily occluded, the per-frame reconstruction cannot recover the articulation, and the physics refiner has no reliable reference to track.

Training cost. The residual reinforcement learning stage is inefficient and slow to train. This cost is the main bottleneck for scaling Art2Sim to larger and more diverse interaction data.

Embodiment gap. The G1 humanoid has large hands that struggle to manipulate fine handle gaps. This limits direct deployment of the recovered references on real hardware for thin or recessed handles.

Asset requirement. Each input video must be paired with an articulated asset that provides geometry and kinematics. Joint asset retrieval or recovery could extend Art2Sim to videos without an available CAD model.

D.2 Future Work

We plan to address these limitations along several directions. First, we aim to scale Art2Sim to objects with multiple active joints and to scenes with multiple interacting objects, where the single-object assumption no longer holds. Second, we aim to close the embodiment gap with finer hands or more precise contact modeling so that the recovered references transfer to real dexterous hardware. Third, we aim to accelerate the residual reinforcement learning stage, for example with distilled or amortized policies, to lower the cost of scaling to larger datasets. Finally, we aim to strengthen contact perception and reference uncertainty modeling and to move toward real-world deployment on humanoid robots.

D.3 Broader Impact

Art2Sim generates physically executable full-body interactions with articulated objects. On the beneficial side, it can support digital human animation, embodied AI research, and the creation of demonstration data for robot manipulation without expensive motion capture or manual annotation. On the risk side, generated human interaction could be misused to fabricate or misrepresent human behavior in video. The method also inherits privacy concerns from the human videos it consumes, and its learned policies reflect the biases present in the training data and the simulation environment. We therefore encourage using Art2Sim only with appropriately licensed or consented footage and applying watermarking or provenance checks to any synthesized interaction that is released publicly.

E AI Use Statement

In this work, we used generative AI tools to help develop theoretical models and conceptual frameworks, propose and refine hypotheses, design and provide feedback on research methodology and experiments, and implement methods. We did not use generative AI tools to formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in writing proofs, assist with translation, support qualitative or thematic data analysis, or interpret results. The generation of synthetic datasets and the cleaning or reformatting of datasets were not applicable to this work. Additionally, we used generative AI tools to create or modify scientific figures or images, summarize or analyze existing literature, brainstorm, and source or search for information.

AI assistance was used as a supporting tool during ideation, literature search, experimental planning, implementation, and figure preparation. The authors reviewed and verified all AI-assisted outputs.

In particular, ideas and literature suggestions were checked against the relevant papers, code suggestions were adapted and tested in the project implementation, and figures and visual assets were inspected for correctness and consistency with the reported experiments. No AI-generated output was accepted without human review, and the authors take full responsibility for the final content of this work, including all text, claims, code, figures, and other artifacts produced with the aid of generative AI.