Counterfactual Video Generation for Humanoid Loco-Manipulation

Zihan Wang[1], Zhen Wu[1], Pieter Abbeel[1,2†], Rocky Duan[1†], Jitendra Malik[1,2†],
Carmelo Sferrazza[1†], Karen Liu[1,4†], Guanya Shi[1,3†], Angjoo Kanazawa[1,2†]
[1]Amazon Frontier AI & Robotics (FAR) [2]University of California, Berkeley
[3]Carnegie Mellon University [4]Stanford University
[†]Amazon FAR team co-leads.

***Abstract*—Teaching humanoids loco-manipulation skills, such as carrying diverse objects, via visual imitation is a promising path toward generalist robots. However, acquiring high-quality, full-body human videos that span a wide variety of object interactions is difficult, leaving too few demonstrations to learn a generalist policy. We propose PRISM, a real-to-sim-to-real framework that overcomes this limitation by amplifying a handful of real videos into a large, diverse training set. PRISM first generates hundreds of diverse "counterfactual" human–object interaction videos via video-to-video (V2V) generation from a few exemplar real videos. Our contact-anchored real-to-sim pipeline then reconstructs both human and object motions, retargeting this imperfect video data into physically plausible trajectories. The intra-class variability across these counterfactual videos lets us train a single policy that generalizes to unseen objects within each category. We demonstrate the full pipeline by deploying this policy on a real robot without any real-world fine-tuning. Using only onboard depth observations, our humanoid picks up, carries, and drops objects—including boxes, barrels, bins, and balls—across novel instances, sizes, and initial configurations.**

I. Introduction

How do we teach humanoids to interact with the objects around us? A promising avenue is visual imitation learning: observing human demonstrations to understand how to approach, lift, and carry various objects. Recent advances in real-to-sim-to-real pipelines [1, 30] have shown that humanoids can acquire contextual whole-body locomotion skills directly from human videos, bypassing the need for hand-tuned reward engineering. This paradigm suggests a scalable path toward generalist robots by tapping into internet-scale video data.

However, scaling visual imitation faces an unspoken data filtering challenge. While the internet certainly contains videos useful for robot learning, finding high-quality, full-body clips of humans carrying everyday objects—properly framed and relatively unoccluded—is practically difficult. Because these interactions represent "boring" data for human viewers, they are not easily searchable and require extensive filtering mechanisms to isolate. Yet, this data does exist implicitly in modern video generative models [2, 37], which act as highly compressed representations of the internet, capturing dense intuitive physics and structural variations.

In this work, we propose *PRISM*, a framework that utilizes video-to-video (V2V) generation to extract this latent knowledge, synthesizing a vast variety of interaction clips from a small set of real exemplar videos. We call these generated clips "counterfactual videos," as they depict how the underlying human motion naturally adapts when the handled objects change in size, shape, and initial configuration. Given these videos, we develop a real-to-sim pipeline that reconstructs human motion, the static scene, and dynamic object geometry and motion from monocular video in a unified world frame. These reconstructed interactions are then retargeted to a humanoid and used to train a depth-based generalist policy. Specifically, starting from only 8 real-world seed videos of humans carrying boxes, we generate 256 counterfactual human–object interaction videos by replacing the carried object with diverse categories, including boxes, balls, bins, and barrels, while varying object scale and initial configuration. Training on these *PRISM* reconstructions enables a single policy to pick up, carry, and drop unseen real-world instances from the same categories, and further transfer to out-of-distribution object categories not present in the counterfactual videos, as shown in Figure **??**.

The core technical challenge is to turn ambiguous monocular counterfactual videos into physically plausible robot demonstrations. Our key insight is not to reconstruct the human, object, and scene independently, but to carry sparse contact anchors through the entire real-to-sim-to-real pipeline as task-level invariants. These anchors are defined in the object frame and encode where the interaction should be maintained, rather than merely when contact occurs. In reconstruction, they couple human motion, object motion, and scene geometry, enabling contact-consistent object trajectories even under occlusion and imperfect generation. In retargeting, the same anchors specify the intended robot end-effector contacts, preserving the manipulation structure despite differences between human and humanoid morphology. This contact-anchored representation turns noisy monocular reconstructions into simulation-ready demonstrations, which are used to train a privileged policy that co-tracks robot-object pose and then distilled into a depth-based generalist policy with joystick control.

We summarize our contributions as follows: (1) We introduce counterfactual video generation as a scalable data augmentation paradigm for human–object interaction learning. (2) We propose a contact-anchored real-to-sim and retargeting pipeline that turns monocular reconstructions into physically plausible robot–object demonstrations. (3) We demonstrate the first real-to-sim-to-real pipeline that trains a unified whole-

body humanoid policy that generalizes across diverse objects and enables joystick-controlled pick-up, carry, and drop behaviors from onboard depth observations.

II. Related Work

a) Humanoid Loco-Manipulation.: Humanoid loco-manipulation has been widely studied in both computer graphics [15, 32, 34, 10, 11] and robotics [31, 38, 36], where the goal is to coordinate whole-body locomotion and object interaction simultaneously. Some humanoid systems [31, 38] learn interaction skills by tracking human motion references, but require dense reference motions or object trajectories during inference, limiting generalization to unseen objects and interaction configurations. More recent methods [13, 8, 24] reduce this dependency by learning more autonomous interaction policies from sparse goals or interaction-centric representations. Similarly, PRISM trains a unified whole-body humanoid policy that enables zero-shot joystick-controlled pick-up, carry, and drop behaviors from onboard depth observations, without reference motions during deployment.

b) Learning from Human Video.: Recent works have leveraged human videos as scalable demonstrations to transfer skills to humanoid robots [7, 31, 14]. This paradigm has been further extended by jointly reconstructing human motion and scene geometry, enabling terrain-aware humanoid skills [1]. More recent methods also model dynamic object interactions, but still rely on nontrivial manual effort, such as contact annotations or object-specific interaction priors [31, 25]. In contrast, PRISM automatically infers contact signals from monocular videos to jointly regularize human motion, object dynamics, and scene geometry within a unified real-to-sim pipeline.

c) Video Model as Data++.: Recent works have explored video generation for robot learning under two main paradigms. One line uses video models as inference-time planners or action generators, synthesizing future visual rollouts to guide closed-loop manipulation [4, 5, 2, 37]. While such methods provide flexible visual reasoning, they require expensive test-time generation and may suffer from an executability gap between plausible videos and physically feasible robot actions. Another line uses video models offline to synthesize robot demonstrations or augment robot datasets before policy learning [16, 9]. PRISM follows the offline-generation paradigm, but differs in generating counterfactual human interaction videos through grounded V2V generation rather than directly generating robot videos from text or images. By conditioning on real demonstrations, V2V preserves realistic motion, lighting, and affordance-aware contacts, while a simple unified prompt can scale one exemplar into diverse object categories, poses, and intra-class variations.

III. Real-to-Sim Data Acquisition

Given a monocular video depicting a human interacting with a static scene and a dynamic object, our goal is to recover camera parameters $(K, \{T_i^c\})$, a static scene mesh M_s, a dynamic object with its metric mesh M_o and its per-frame poses $\{T_i^o\}$, and 4D SMPL-H [17] motion, all in a unified world frame. Then we retarget the recovered data to humanoid–object trajectories for policy learning.

A. Counterfactual Interaction Video Generation

Scaling real-to-sim learning with online videos is challenging: useful full-body human–object interaction clips are difficult to retrieve, often contain complex camera motion or occlusions, and rarely cover the object and behavior variations needed for robot learning. We instead use video-to-video generation to turn a small set of real interaction videos into diverse *counterfactual interaction videos*. Unlike text-to-video generation, V2V generation remains grounded in a real sequence, preserving the realistic layout, lighting and coarse task structure while editing the manipulated object.

Importantly, the counterfactual variation is not limited to object geometry or appearance. When the object category, size, pose, or placement changes, the object's manipulation affordances also change, and the human behavior adapts accordingly: the generated person may bend lower, adjust hand spacing, adapt the contact strategy based on the object's affordances, or carry the object differently. Thus each generated clip provides an object-conditioned interaction strategy, rather than merely pairing the same human motion with a new object. This allows PRISM to obtain behavior-level variations in interaction data that are difficult to hard-code through geometry-level augmentation alone. Learning such adaptations from scratch would require exhaustive task-specific RL reward design; PRISM instead uses the video model as an offline prior over plausible human adaptations.

In practice, we use real-world seed video from BEHAVE dataset [3] for V2V generation. Given a seed video, we prompt SeedDance 2.0 [20] to replace the manipulated object with a target category while preserving the original background, lighting, camera viewpoint, and temporal continuity (details in Appendix. A). We instantiate the prompt with four target categories: boxes, bins, barrels, and balls. Rather than specifying instance geometry, we use category-level prompts, allowing the video model to sample intra-class variation in object shape, size, appearance, pose, and placement, together with the corresponding human interaction behavior. For each seed video, we generate 8 samples per category, yielding $N = 32$ counterfactual videos (CF-videos) per seed. Through this counterfactual paradigm, a small set of real-world videos can be scaled into hundreds of high-quality video clips. These generated videos are then processed by our real-to-sim pipeline below.

B. Contact-Anchored Real-to-Sim

a) Background.: Previous work CRISP [27] integrates an HMR network with visual SLAM to reconstruct a metrically-consistent human–scene–camera representation from monocular video, including camera intrinsics $K \in \mathbb{R}^{3\times3}$, per-frame camera poses $T_i = [R_i \mid t_i] \in SE(3)$, a metric-scale scene point cloud $\tilde{P}$, and temporally aligned 4D human motion in

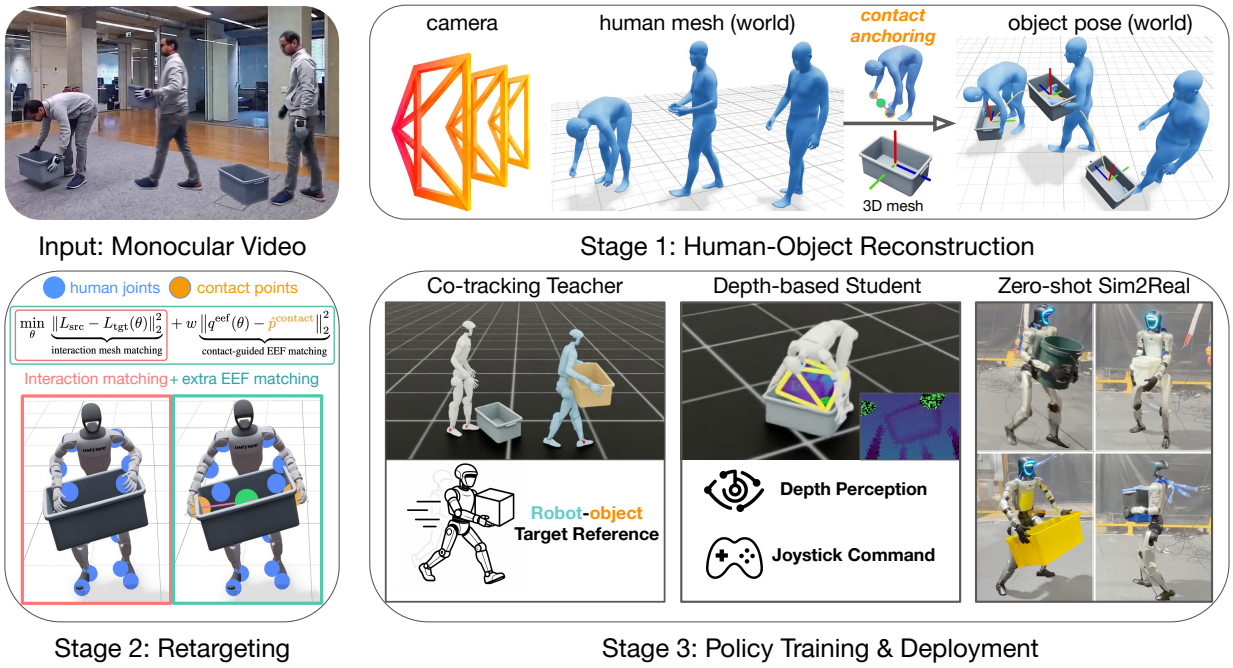

Fig. 1. **PRISM overview.** PRISM turns monocular human–object videos into deployable humanoid loco-manipulation skills. It reconstructs the camera, human motion, object geometry, and 6D object motion in a shared world frame, using object-frame contact anchors to constrain object pose optimization under monocular ambiguity. The same anchors are used in retargeting to preserve interaction phases and match robot end-effectors to intended contact points. The retargeted demonstrations train a privileged co-tracking teacher, which is distilled into a depth-based student policy conditioned on onboard depth and joystick commands for zero-shot sim-to-real deployment.

a unified world coordinate frame. We use it as backend and extend it to reconstruct both the geometry and motion of dynamic objects.

b) Object Geometry and Motion Reconstruction.: Given dynamic object masks extracted by SAM 2 [18], we reconstruct the object geometry using SAM3D [21], together with the camera poses and metric depth estimated by CRISP. This produces the object mesh $\mathcal{M}o$ and its initial world pose $To^{t=0} \in SE(3)$. We avoid visual 6D trackers such as FoundationPose [29], as hand/body occlusions often cause drift, scale errors, and pose jumps in monocular videos (Table III). PRISM instead uses the human motion and contact phase to recover a task-structured object trajectory. For the pick–carry–drop interactions considered in this work, the object follows a simple hybrid structure: it is supported by the scene outside contact, and moves with the human during stable contact. Thus, outside contact intervals, we keep the object pose fixed and grounded by scene geometry support. In contrast to prior work that relies on manual contact annotations [30, 25], we detect contact from the 2D human and object masks by measuring their 2D placement across frames. For each detected contact phase $[t_1, t_2]$, we use the SMPL-X palm as the human-body anchor and denote its world-frame pose as T^t_{palm}. We then express the object pose in this palm frame at the beginning of the interval and propagate it by keeping the palm-relative transform fixed:

$$\begin{aligned} T^t_o &= T^t_{\text{palm}} T_{\text{palm}\to o} = T^t_{\text{palm}} \left(T^{t_1}_{\text{palm}}\right)^{-1} T^{t_1}_o, \\ T_{\text{palm}\to o} &= \left(T^{t_1}_{\text{palm}}\right)^{-1} T^{t_1}_o, \qquad t \in [t_1, t_2]. \end{aligned} \tag{1}$$

where T^t_o is the world-frame object pose at frame t, T^t_{palm} is the world-frame pose of the selected SMPL-X palm anchor, and $T_{\text{palm}\to o}$ is the fixed palm-relative object transform. This formulation anchors object motion to the SMPL-X palm along with human motion during contact intervals.

c) Contact-Anchored Retargeting.: Prior retargeting methods [35] preserve human–object relations but assume clean, physically plausible inputs. Monocular reconstructions often violate this assumption (pink in Fig. 1), causing retargeting to preserve the reconstruction errors. We therefore propose contact-anchored retargeting, which treats the reconstructed full-body human motion as a kinematic reference while using object-frame contact anchors as the reliable interaction interface. These anchors specify where the robot end-effectors should act on the object, allowing the retargeting solver to correct noisy reconstructions into physically plausible robot–object motions.

For each contact phase, we derive contact anchors from the reconstructed human–object geometry in Stage 1. Specifically, we intersect the line segment connecting the SMPL-X left and right palm centers with the object mesh. Following interaction-preserving retargeting [35], we keep the constrained IK and add a contact-anchor term to the original objective. At each

frame, we solve:

$$\begin{aligned} dq_a^\star &= \underset{dq_a \in \mathcal{C}(q_a)}{\arg\min}\ E_{\text{base}}(q_a, dq_a) + w_c \|\Delta x_{\text{eef}}\|_2^2, \\ \Delta x_{\text{eef}} &= x_{\text{eef}}(q_a) + J_{\text{eef}}(q_a) dq_a - c. \end{aligned} \tag{2}$$

Here E_{base} denotes the interaction mesh matching objective introduced from [35], c is the target contact point in object coordinate estimated from the human–object reconstruction in Fig. 1, and $x_{\text{eef}}(q_a) + J_{\text{eef}}(q_a) dq_a$ is the linearized end-effector position. We update $q_a \leftarrow q_a + dq_a^\star$ at each frame and warm-start the next frame from the updated configuration.

One might attribute these artifacts to the reconstruction pipeline. We take a different view: imperfect reconstruction is unavoidable in monocular real-to-sim stage, as no existing pipeline can recover physically plausible human–object motions. Our contact-anchored real-to-sim system closes this gap by optimizing sufficiently close reconstructions into physically plausible robot demonstrations.

IV. Learning a Unified Visuomotor Interaction Policy

Using the reconstructed robot-object trajectories (details are in Appendix A), our goal is to train a unified humanoid policy capable of pick-up, carry, and drop behaviors across diverse objects. The policy receives onboard depth observations and joystick commands, and autonomously performs object interaction behaviors conditioned on the perceived object geometry and placement. Similar to prior visuomotor systems [33, 6], we adopt a two-stage training framework. We first train privileged co-tracking teacher policies in simulation using full-state observations. We then distill these teachers into a single depth-based student policy using a combination of DAgger [19] and reinforcement learning (RL), enabling zero-shot sim-to-real deployment. An overview is shown in Fig. 1.

A. Training Co-Tracking Teacher Policies

We formulate humanoid object interaction as a co-tracking problem, where the policy simultaneously tracks both the humanoid motion and the dynamic object trajectory reconstructed from monocular videos. We follow prior works [12, 35] for motion-tracking training and refer readers to these works for implementation details.

Observations. Our teacher observations include reference joint positions and velocities, reference pelvis pose error, pelvis linear and angular velocity, joint positions and velocities, and the previous action. We additionally include the current object pose in the robot frame, the target object pose, and the object's 3D bounding-box dimensions, making the policy explicitly aware of both the desired robot motion and the object state it must manipulate.

Rewards, Terminations, and Domain Randomization. The reward primarily consists of robot pose tracking, object pose tracking, action rate, joint limits, and collision penalties. We additionally introduce a contact-aware interaction reward using the recovered human-object contact anchors. Specifically, we encourage the robot end-effectors to reach the reconstructed contact locations and apply sufficient contact force:

$$\begin{aligned} R_{\text{contact},i} = &\ \exp\left(-\frac{\|\mathbf{p}_{\text{eef},i} - \mathbf{p}_{\text{target},i}\|_2}{\sigma_{\text{pos}}}\right) \\ &\cdot \min\left(\exp\left(\frac{\|\mathbf{F}_{\text{contact},i}\|_2 - \mathbf{F}_{\text{thres}}}{\sigma_{\text{frc}}}\right), 1\right). \end{aligned} \tag{3}$$

where $\mathbf{p}_{\text{eef},i}$ denotes the end-effector position, $\mathbf{p}_{\text{target},i}$ is the reconstructed target contact point, and $\mathbf{F}_{\text{contact},i}$ is the contact force magnitude. This contact-aware reward stabilizes grasping and carry behaviors under noisy monocular reconstructions and substantially improves interaction quality during policy learning. We also adopt early termination and domain randomization following [35].

B. Distilling a Unified Depth-Based Student Policy

While privileged teachers can effectively learn stable object interaction behaviors in simulation, they rely on information unavailable on real hardware. We therefore distill the teachers into a deployable depth-based student policy using onboard perception only.

DAgger + RL Distillation. Following [33], we distill the teacher policies using a combination of DAgger and PPO objectives:

$$\mathcal{L} = \lambda \mathcal{L}_{\text{PPO}} + (1 - \lambda) \mathcal{L}_D, \tag{4}$$

where $\mathcal{L}_D$ denotes the DAgger imitation loss and $\mathcal{L}_{\text{PPO}}$ is the RL objective. We follow the same curriculum as [33], gradually increasing λ throughout training.

Observations. The student policy receives proprioception, joystick commands, and onboard depth observations. Proprioception includes the pelvis gravity vector and angular velocity, joint positions and velocities, and the previous action. The joystick command consists of a sparse root command $c_t^{\text{js}} = [\Delta x_t, \Delta y_t, \Delta \text{yaw}_t]$, which specifies the desired relative root motion in the robot root-heading frame, and a binary drop command $b_t^{\text{drop}} \in 0, 1$. During training, c_t^{js} is computed from the relative reference root pose in the current root-heading frame, while b_t^{drop} is set to 0 before the carry-end time and 1 thereafter. For depth sensing, we additionally randomize the camera pose and inject depth artifacts, including multiplicative depth noise, dropout, holes, edge artifacts, Gaussian noise, and depth offsets. The critic receive the same privileged observation as teacher.

Curriculum. We use an object-centric curriculum that anneals training from box-only interactions to the full object mixture. Boxes provide simple geometry and stable contacts for early learning, while later stages introduce objects with more diverse shapes, scales, and interaction modes. We empirically find that this design stabilizes policy learning during student distillation.

Training Details. Both teacher and student policies are 3-layer MLPs trained for 20K iterations with 4096 parallel environments on 8 NVIDIA L40S GPUs. During distillation, we use the teacher rollout instead of original kinematic motion. For more details refer to Appendix. A.

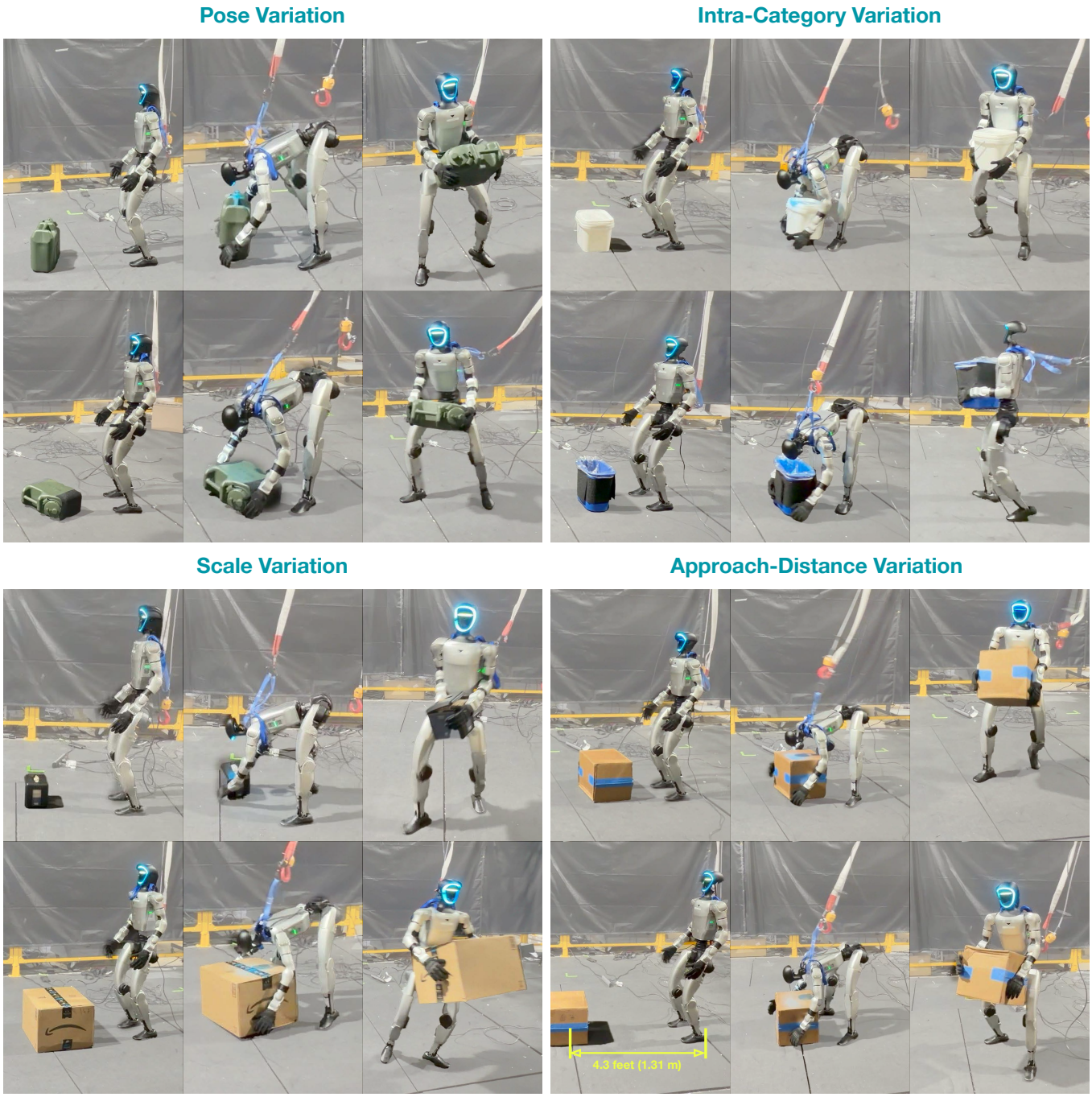

Fig. 2. **Policy rollout under real-world object variations.** Our single unified policy zero-shot transfers to the real robot across object-pose, intra-category, scale, and approach-distance variations. The policy is agnostic to object placement and initial pose (*top-left*), handles objects with different topology and appearance within the same category (*top-right*), and generalizes to large scale variations (*bottom-left*). It further adapts to object distance using onboard depth, either approaching (up to 4.3 feet!) before grasping or directly grasping when the object is within reach (*bottom-right*).

V. Results

Overview. We demonstrate that humanoid robots can learn skills that generalize to diverse objects by imitating pure counterfactual generated videos. We first evaluate the effectiveness of our reconstructed data by comparing with the policy trained with OMOMO [10] data only. Next, we evaluate the real-world performance over 20+ diverse objects. Finally, we conduct necessary ablation study to validate the necessity of each module in our framework design. All evaluations are conducted in MuJoCo [22] under a sim-to-sim setting, except for Sec. V-B reports real-world robot experiments.

A. Reconstruction Quality from Real-to-Sim

Evaluation Data. We process and filter OMOMO to retain 63 high-quality human–object interaction sequences, then double the data by adding an object-scaled variant for each sequence using OmniRetarget [35]. This gives 126 sequences, with 90 for *OMOMO-Train* and 36 held out as *OMOMO-Test*. For PRISM, we use 80 training reconstructions as *PRISM-ID* and evaluate 48 sequences *PRISM-OOD* on unseen reconstructed interactions with chairs, tables, lamps, and monitors.

Evaluation Metrics. For the distilled student policy, we derive joystick commands from the reference motion and render depth observations in simulation to drive the policy. A rollout is successful if the robot completes pick–carry–drop and places the object at the desired target position.

Discussion. OMOMO-trained policies perform well across OMOMO sequences, but transfer poorly to PRISM interactions, especially with out-of-domain objects. We identify three main failure modes: the policy struggles with unseen depth observations, overfits to close-range interactions and attempts to grasp before approaching distant objects, and fails to stabilize carry on novel object geometries. These failures suggest that OMOMO alone lacks the variation in perception, object configuration, and interaction dynamics required for robust loco-manipulation. In contrast, PRISM-trained policies generalize across both domains, demonstrating the broader transferability of PRISM data.

TABLE I
CROSS-DOMAIN EVALUATION.

OMOMO-trained policies perform well on held-out OMOMO interactions but fail to transfer to PRISM interactions. PRISM-trained policies generalize across held-out OMOMO and PRISM interactions, including out-of-domain PRISM objects.

Training Data	OMOMO Train	OMOMO Test	PRISM ID	PRISM OOD
OMOMO (90)	94.44%	91.67%	23.75%	12.50%
PRISM ID(80)	**98.89%**	**100%**	**96.25%**	**72.92%**

B. Real-World Deployment

We deploy our controller on a 29-DoF Unitree G1 humanoid and run it onboard at 50 Hz. We apply the PD controller setting following [12]. The robot perceives depth from head-mounted D435i at 30 Hz. We feed joystick targets from a human operator. All operations and policy are run onboard. For real-world evaluation, we define success as completing the full reach–grasp–carry sequence: the robot must reach the object, grasp it, and carry it stably for at least 3 meters without dropping the object or falling. For each object, we place it in front of robot in random pose and distance and try it 5 times to calculate the real-world success rate. An overview of the objects we tested is included in Fig. 4 for reference. We highlight that no scans or reconstructions of the real-world test objects are used for training, showing that our policy can zero-shot generalize across a broad object distribution.

TABLE II
REAL-WORLD EVALUATION.

Each in-domain category contains three object instances; each object is tested with five trials. Object details are shown in Fig. 4.

	In-domain				**Out-of-domain**							
Metric	Ball	Bin	Barrel	Box	Paper towel	Helmet	Table	Backpack	Lamp	Chair	Kettle	Chick toy
Success / Trials	12 / 15	14 / 15	14 / 15	15 / 15	5 / 5	3 / 5	5 / 5	5 / 5	3 / 5	4 / 5	4 / 5	3 / 5
Success Rate	80%	93%	93%	100%	100%	60%	100%	100%	60%	80%	80%	60%

x

TABLE III
ABLATION STUDY.

All rows are evaluated using PRISM reconstructed data.

	Contact-anchored Pose	Contact-anchored Retargeting	Contact Reward	Suc. (ID)	Suc. (OOD)
Baseline	✗	✗	✗	22.50%	12.50%
+ anchored object pose	✓	✗	✗	58.75%	29.17%
+ contact-aware retargeting	✓	✓	✗	86.25%	68.75%
+ contact reward (PRISM)	✓	✓	✓	96.25%	72.92%

Using only the 8 original seed videos leads to in-domain overfitting and poor generalization. We therefore use V2V counterfactual generation for data expansion and ablate the key real-to-sim-to-real components on this expanded dataset. Following [28], a good reconstruction for robot learning should be physically plausible, simulatable and useful for policy training. Therefore, we use the reconstructed trajectories to train policies in simulation, and report the downstream task success rate as the main metric. We define the baseline as: initialize object geometry and poses by SAM3D [21], then track it by FoundationPose [29], and retarget it into simulation-ready robot–object demonstrations with OmniRetarget [35]. Table III shows that better real-to-sim data directly benefits downstream policy learning. The baseline produces

inaccurate robot–object interactions that limits effective policy learning. Anchored object pose and contact-aware retargeting progressively improve the data quality thus lead to stronger downstream performance. The contact reward further improves the policy by encouraging stable, task-relevant robot–object contacts in training.

VI. Conclusion

We introduce PRISM, a real-to-sim-to-real framework that converts generated monocular human–object videos into simulation-ready humanoid demonstrations. By using contact as a geometric and physical anchor, PRISM corrects noisy reconstructions and produces feasible robot–object trajectories for policy learning. The learned unified policy operates from onboard depth and generalizes across diverse objects for pick-up, carry, and drop, demonstrating generated videos as a scalable data source for humanoid loco-manipulation.

VII. Limitations and Future Work

Our pipeline delivers encouraging real-world results, yet several practical weaknesses remains.

Reconstruction and Retargeting. Monocular 4D human–object–scene recovery remains challenging, with most failures from inaccurate interaction modeling. Training a unified framework that can jointly reconstruct humans, objects, scenes, and contacts is an important future direction.

Data Scale. Although our results provide initial evidence for the effectiveness of the real-to-sim-to-real framework, the scale of our study remains limited. We evaluate only $N = 32$ examples with 256 human-object reconstructions, which is insufficient to fully characterize scaling behavior.

Object Physics. Our training does not model class-specific object physics; instead, object categories share the same material, friction, mass, and inertia settings. This simplification can lead to inaccurate contact dynamics, especially for objects with smooth surfaces or non-uniform mass distributions. For example, the helmet failure pattern is to slip out of the hands despite accurate grasping.

Simulation Limitations. We model dynamic objects as rigid bodies. Thus the policy can struggle with articulated or deformable objects whose contact geometry may change during interaction. For example, foldable chairs may shift through internal joints, causing unstable grasps or loss of control.

References

[1] Arthur Allshire, Hongsuk Choi, Junyi Zhang, David McAllister, Anthony Zhang, Chung Min Kim, Trevor Darrell, Pieter Abbeel, Jitendra Malik, and Angjoo Kanazawa. Visual imitation enables contextual humanoid control. *arXiv preprint arXiv:2505.03729*, 2025.

[2] Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. *arXiv preprint arXiv:2409.16283*, 2024.

[3] Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: Dataset and method for tracking human object interactions. In *CVPR*, 2022.

[4] Qingwen Bu, Jia Zeng, Li Chen, Yanchao Yang, Guyue Zhou, Junchi Yan, Ping Luo, Heming Cui, Yi Ma, and Hongyang Li. Closed-loop visuomotor control with generative expectation for robotic manipulation. *Advances in Neural Information Processing Systems*, 37:139002–139029, 2024.

[5] Boyuan Chen, Tianyuan Zhang, Haoran Geng, Caiyi Zhang, Peihao Li, Kiwhan Song, William T Freeman, Jitendra Malik, Pieter Abbeel, Russ Tedrake, et al. Large video planner enables generalizable robot control. *arXiv preprint arXiv:2512.15840*, 2025.

[6] Xuxin Cheng, Kexin Shi, Ananye Agarwal, and Deepak Pathak. Extreme parkour with legged robots. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11443–11450. IEEE, 2024.

[7] Tairan He, Jiawei Gao, Wenli Xiao, Yuanhang Zhang, Zi Wang, Jiashun Wang, Zhengyi Luo, Guanqi He, Nikhil Sobanbab, Chaoyi Pan, et al. Asap: Aligning simulation and real-world physics for learning agile humanoid whole-body skills. *arXiv preprint arXiv:2502.01143*, 2025.

[8] Xialin He, Sirui Xu, Xinyao Li, Runpei Dong, Liuyu Bian, Yu-Xiong Wang, and Liang-Yan Gui. Ultra: Unified multimodal control for autonomous humanoid whole-body loco-manipulation. *arXiv preprint arXiv:2603.03279*, 2026.

[9] Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, et al. Dreamgen: Unlocking generalization in robot learning through video world models. *arXiv preprint arXiv:2505.12705*, 2025.

[10] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023.

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

[12] Qiayuan Liao, Takara E Truong, Xiaoyu Huang, Yuman Gao, Guy Tevet, Koushil Sreenath, and C Karen Liu. Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion. *arXiv preprint arXiv:2508.08241*, 2025.

[13] Yutang Lin, Jieming Cui, Yixuan Li, Baoxiong Jia, Yixin Zhu, and Siyuan Huang. Lessmimic: Long-horizon humanoid interaction with unified distance field representations. *arXiv preprint arXiv:2602.21723*, 2026.

[14] Jiageng Mao, Siheng Zhao, Siqi Song, Tianheng Shi, Junjie Ye, Mingtong Zhang, Haoran Geng, Jitendra Malik, Vitor Guizilini, and Yue Wang. Learning from massive human videos for universal humanoid pose control. *arXiv preprint arXiv:2412.14172*, 2024.

[15] Liang Pan, Zeshi Yang, Zhiyang Dou, Wenjia Wang, Buzhen Huang, Bo Dai, Taku Komura, and Jingbo Wang. Tokenhsi: Unified synthesis of physical human-scene interactions through task tokenization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5379–5391, 2025.

[16] Shivansh Patel, Shraddhaa Mohan, Hanlin Mai, Unnat Jain, Svetlana Lazebnik, and Yunzhu Li. Robotic manipulation by imitating generated videos without physical demonstrations. *arXiv preprint arXiv:2507.00990*, 2025.

[17] 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.

[18] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe 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.

[19] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011.

[20] Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

[21] 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. *arXiv preprint arXiv:2511.16624*, 2025. URL https://arxiv.org/abs/2511.16624.

[22] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco:

A physics engine for model-based control. In *IROS*, 2012.

[23] Zifu Wan, Yaqi Xie, Ce Zhang, Zhiqiu Lin, Zihan Wang, Simon Stepputtis, Deva Ramanan, and Katia P Sycara. Instructpart: Task-oriented part segmentation with instruction reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24202–24227, 2025.

[24] Huayi Wang, Wentao Zhang, Runyi Yu, Tao Huang, Junli Ren, Feiyu Jia, Zirui Wang, Xiaojie Niu, Xiao Chen, Jiahe Chen, Qifeng Chen, Jingbo Wang, and Jiangmiao Pang. Physhsi: Towards a real-world generalizable and natural humanoid-scene interaction system. *arXiv preprint arXiv:2510.11072*, 2025.

[25] Yinhuai Wang, Qihan Zhao, Yuen Fui Lau, Runyi Yu, Hok Wai Tsui, Qifeng Chen, Jingbo Wang, Jiangmiao Pang, and Ping Tan. Humanx: Toward agile and generalizable humanoid interaction skills from human videos. *arXiv preprint arXiv:2602.02473*, 2026.

[26] Zihan Wang, Jeff Tan, Tarasha Khurana, Neehar Peri, and Deva Ramanan. Monofusion: Sparse-view 4d reconstruction via monocular fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8252–8263, 2025.

[27] Zihan Wang, Jiashun Wang, Jeff Tan, Yiwen Zhao, Jessica Hodgins, Shubham Tulsiani, and Deva Ramanan. Crisp: Contact-guided real2sim from monocular video with planar scene primitives. *arXiv preprint arXiv:2512.14696*, 2025.

[28] Zihan Wang, Jiashun Wang, Jeff Tan, Yiwen Zhao, Jessica K Hodgins, Shubham Tulsiani, and Deva Ramanan. Contact-guided real2sim from monocular video with planar scene primitives. In *The Fourteenth International Conference on Learning Representations*, 2026.

[29] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects, 2024. URL https://arxiv.org/abs/2312.08344.

[30] 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.

[31] 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.

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

[33] Zhen Wu, Xiaoyu Huang, Lujie Yang, Yuanhang Zhang, Koushil Sreenath, Xi Chen, Pieter Abbeel, Rocky Duan, Angjoo Kanazawa, Carmelo Sferrazza, et al. Perceptive humanoid parkour: Chaining dynamic human skills via motion matching. *arXiv preprint arXiv:2602.15827*, 2026.

[34] 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 Computer Vision and Pattern Recognition Conference*, pages 12266–12277, 2025.

[35] Lujie Yang, Xiaoyu Huang, Zhen Wu, Angjoo Kanazawa, Pieter Abbeel, Carmelo Sferrazza, C Karen Liu, Rocky Duan, and Guanya Shi. Omniretarget: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction. *arXiv preprint arXiv:2509.26633*, 2025.

[36] Lujie Yang, Xiaoyu Huang, Zhen Wu, Angjoo Kanazawa, Pieter Abbeel, Carmelo Sferrazza, C. Karen Liu, Rocky Duan, and Guanya Shi. Omniretarget: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction. In *ICRA*, 2025.

[37] Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026.

[38] 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.

APPENDIX

Appendix A: Counterfactual Video Generation Details We use the SeedDance 2.0 web interface to perform video-to-video generation. Given each seed video, we instantiate the counterfactual prompt with each object category and generate eight samples per category (represented as <CLS>). This lightweight prompting procedure turns a small number of seed videos into a scalable and diverse set of counterfactual human–object interaction videos. Instead of relying on multi-view captures [26] or affordance learning [23], our policy learns adaptive skill in handling diverse real-world objects.

Prompt template.

```
<VIDEO>

In a real-world continuous footage.
Preserve the reference video's original
background, lighting and camera
viewpoint.

replace the box with <CLS>, pick it
up and carry with two hands.
```

Fig. 3. **Video-to-video prompt template.** The input video is represented by <VIDEO>. The highlighted span is the counterfactual instruction to edit the object and motion.

Appendix B: Real-World Test Objects

Appendix C: PRISM training and evaluation data Starting from 256 reconstructed videos, our retargeting pipeline produces 133 feasible robot–object trajectories; the remaining videos are filtered out because they fail to satisfy the constrained solver, mainly due to undesired collisions. We use these trajectories to train a privileged teacher policy that jointly tracks the robot and object motions. At 20K training iterations, the teacher can successfully roll out 80 trajectories, which are then used as demonstrations for student distillation.

In the retargeting stage, most failures are caused by severe collisions. In particular, some target objects are too large relative to the Unitree G1 humanoid, making them physically difficult or even infeasible for the robot to carry. These cases often lead to substantial object–body collisions and eventually result in failed retargeting. Nevertheless, we would like to emphasize that scaling the amount of source data remains a highly promising direction, even under the current conversion rate, as it can still substantially increase the number and diversity of successful retargeted trajectories.

At the time of writing, the teacher policy used for generating demonstrations had not fully converged. We later found that extending teacher-policy training to 40K iterations enabled successful rollouts for 117 trajectories. However, due to time constraints, we were unable to use these improved teacher rollouts for the subsequent distillation stage in the current version of the work.

We additionally reconstruct 48 out-of-domain object instances, consisting of 12 instances each from four categories: tables, chairs, lamps, and monitors. These reconstructed objects are used as out-of-distribution data evaluation.

Appendix D: Training and Distillation Hyperparameters

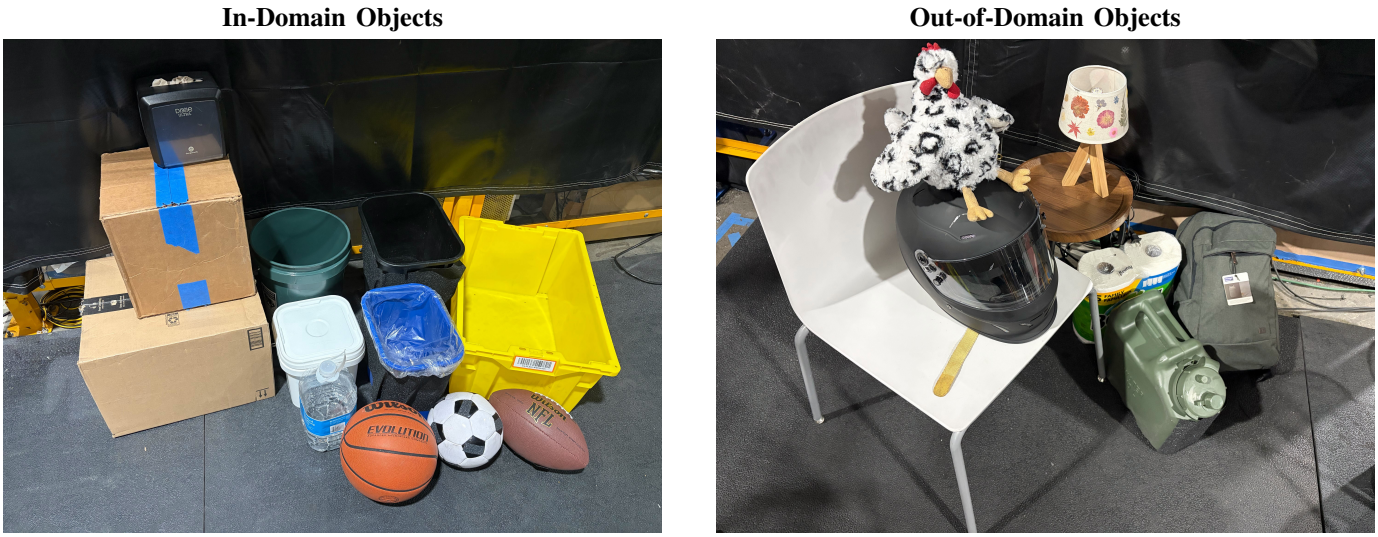

Fig. 4. **Real-world test objects.** We evaluate the policy on both in-domain and out-of-domain objects, spanning substantial variations in appearance, geometry, weights and scale.

TABLE IV
OBSERVATION SPACES.

T/S denote teacher/student, and A/C denote actor/critic. $\checkmark^{\dagger}$ denotes privileged student-critic inputs that are not used by the deployment policy.

Input	**Dim.**	**T-A**	**T-C**	**S-A**	**S-C**
Commands and references					
Motion command m_t	58	✓	✓	–	$\checkmark^{\dagger}$
Reference root position	3	–	✓	–	$\checkmark^{\dagger}$
Reference root orientation	6	✓	✓	–	$\checkmark^{\dagger}$
Tracked body positions	42	–	✓	–	$\checkmark^{\dagger}$
Tracked body orientations	84	–	✓	–	$\checkmark^{\dagger}$
Sparse root command c_t^{js}	3	–	–	✓	–
Drop button b_t^{drop}	1	–	–	✓	–
Robot proprioception and actions					
Base linear velocity	3	–	✓	–	$\checkmark^{\dagger}$
Base angular velocity	3	✓	✓	✓	✓
Joint positions	29	✓	✓	✓	✓
Joint velocities	29	✓	✓	✓	✓
Previous action	29	✓	✓	✓	✓
Proprioceptive history	64	–	–	–	✓
Object state and perception					
Object pose (p_o, R_o)	9	✓	✓	–	$\checkmark^{\dagger}$
Target pose (p_g, R_g)	9	✓	✓	–	$\checkmark^{\dagger}$
Object size	3	✓	✓	–	$\checkmark^{\dagger}$
Object linear velocity	3	–	✓	–	$\checkmark^{\dagger}$
Object angular velocity	3	–	–	–	$\checkmark^{\dagger}$
Depth image / CNN latent	$58\times87 \rightarrow 32$	–	–	✓	–
Total input	–	$175\times5 = 875$	$310\times5 = 1550$	$94 + 32 = 126$	$284 + 64 + 29 = 377$
Deployment	–	train only	train only	✓	train only

TABLE V
MAIN CO-TRACKING REWARD WEIGHTS FOR TEACHER TRAINING.

Tracking terms use exponential kernels with the listed scale parameters.

Reward term	Weight	Parameters
Root position tracking	0.5	$\sigma = 0.3$
Root orientation tracking	0.5	$\sigma = 0.4$
Full-body position tracking	1.0	$\sigma = 0.3$
Full-body orientation tracking	1.0	$\sigma = 0.4$
Body linear velocity tracking	1.0	$\sigma = 1.0$
Body angular velocity tracking	1.0	$\sigma = 3.14$
Object position tracking	1.0	$\sigma = 0.3$
Object orientation tracking	1.0	$\sigma = 0.4$
Offline wrist target guidance	5.0	Object-frame contact points; $\sigma = 0.08$
Offline force-gated contact guidance	10.0	Force threshold 1.0; force $\sigma = 10.0$; contact schedule relaxation 5 steps
Action-rate penalty	-0.1	Squared action difference
Joint-limit penalty	-10.0	Soft joint limit 0.9
Foot/ankle object-contact penalty	-0.5	Threshold 1.0
Lower-body undesired contact penalty	-0.1	Threshold 1.0

TABLE VI
TRAINING AND DISTILLATION HYPERPARAMETERS.

Main hyperparameters for the privileged co-tracking teacher and the depth-based student.

Hyperparameter	**Co-tracking teacher**	**Depth-based student**
Training data	Original motion–object bank	Success-filtered teacher rollouts
Number of clips	133	80
Total environments	8192×8	4096×8
Policy input	Motion command, robot state, object state	Sparse root command, proprioception, buttons, depth
Object input	Current pose, target pose, object size	Current pose through depth and command conditioning
History length	1	1
Actor network	MLP $[512, 256, 128]$	MLP with depth latent input
Critic network	MLP $[512, 256, 128]$	MLP with privileged state
Activation	ELU	ELU
Depth input	None	106×60 warped to 58×87
Depth encoder	None	Small CNN, 32-D latent
Camera range	None	0.3–3.0m
Depth noise	None	Hole prob. 0.2, noise std. 0.03m, offset std. 0.03m
Rollout length	24	24
Minibatches	4	4
Learning epochs	7	5
Optimizer	AdamW	AdamW
Actor learning rate	1.0×10^{-5}	7.0×10^{-5}
Critic learning rate	1.0×10^{-5}	7.0×10^{-5}
Discount factor	$\gamma = 0.99$	$\gamma = 0.99$
GAE parameter	$\lambda = 0.95$	$\lambda = 0.95$
PPO clip range	0.2	0.2
Value loss coefficient	1.0	1.0
Entropy coefficient	0.005	0
Weight decay	10^{-3}	Not used
Max gradient norm	1.0	1.0
Initial policy std.	1.0	0.01
Training iterations	$20,000$	$20,000$
Distillation loss	None	MSE behavior cloning
BC coefficient	None	1.0
DAgger coefficient	None	1.0
PPO coefficient schedule	None	$0.1 \rightarrow 0.9$, step 0.1 every 500 iterations
Teacher-action rollout mix	None	0
Teacher action clip	None	8.0
Contact-aware command	None	Peak-height mode, $\alpha = 0.91$, smoothing 5 steps
Sampling curriculum	Success-rate adaptive clip weighting	Adaptive timestep sampling
Reset schedule	Random episode initialization	Start-at-zero prob. $0.2 \rightarrow 1.0$
Default-pose prepend	0.2s	Not used
Episode length	Motion-dependent	8.0s
Domain randomization	Friction, restitution, COM, joint bias, object mass/inertia, camera noise, pushes	Friction, restitution, COM, joint bias, object mass/inertia, camera noise, pushes
External pushes	Interval $[0.5, 2.0]$s; max velocity $[0.7, 0.7, 0.25, 0.7, 0.7, 1.0]$	Interval $[0.5, 2.0]$s; max velocity $[0.7, 0.7, 0.25, 0.7, 0.7, 1.0]$