TouchTwin: Scaling Human Touch Data via Real-to-Sim-to-Real

Huiwen Han[1,2,*], **Nuo Chen**[2,*], **Ziyao Zeng**[3,*], **Dayou Li**[2], **Hao Wang**[2], **Amir Hossein Rahmati**[2], **Xiaoning Qian**[2], **Bangya Liu**[4], **Zhiwen Fan**[2]

[1]Stanford University, [2]Texas A&M University, [3]Yale University, [4]2077AI
[*]Equal contribution

Human demonstrations are becoming the main source of data for dexterous manipulation, and real-to-sim-to-real pipelines multiply each demonstration in simulation. These pipelines rebuild the interaction from vision, which recovers how the hand moves. Which fingers touch, when, and how hard remain unobserved, because a hand pose shows where each finger stops on the surface rather than how far it is driven to close. This gap is a force gap: on held-out recordings, a hand policy that imitates the observed pose grips with 0.31× the recorded force, and a policy trained on physics replays of vision-tracked poses grips with 1.55×. TouchTwin closes the gap with measured human touch. Each selected TouchScale recording, captured with head and wrist RGB-D cameras and an 880-taxel glove, is rebuilt in MuJoCo. Coding agents write the objects as physics-tested programs, and a per-subject tactile hand replays the interaction with the recorded touch setting the contact and the closing force of each finger. In simulation, a policy trained on touch-grounded replays grips at 0.99× the recorded force and succeeds in 20/20 trials on held-out recordings and 35/36 unseen perturbations; a contact predictor trained on them misses 13% of real contact frames, against 32% for vision-based replays. Each replay also seeds perturbed demonstrations with simulated contact and force labels, rendered as realistic video in new views and scenes.

Date: November 8, 2026
Keywords: Real-to-sim-to-real, Dexterous manipulation, Egocentric data
Project Page: https://touchtwin.github.io

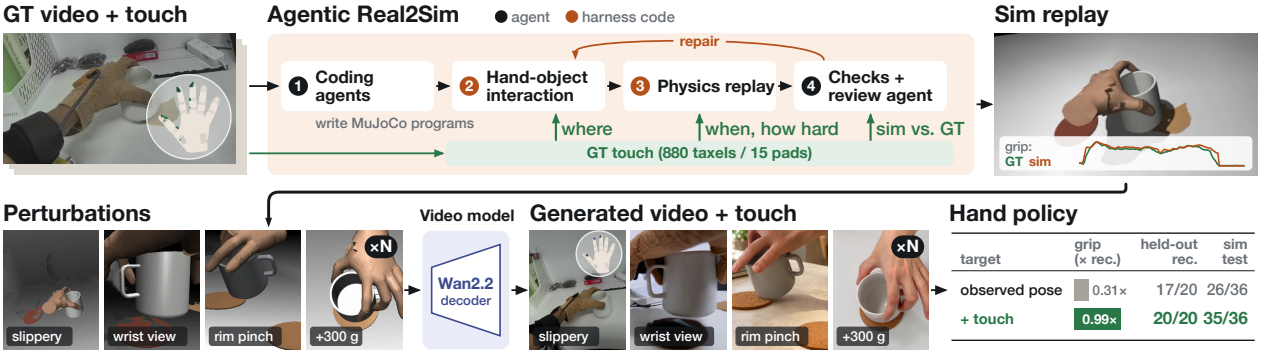


Figure 1 TouchTwin overview. Top: an agentic pipeline rebuilds each recording; GT touch sets where, when and how hard fingers touch and checks the replay. Bottom: each replay yields N perturbed demos, decoded to video in new views and scenes, that train a hand policy.

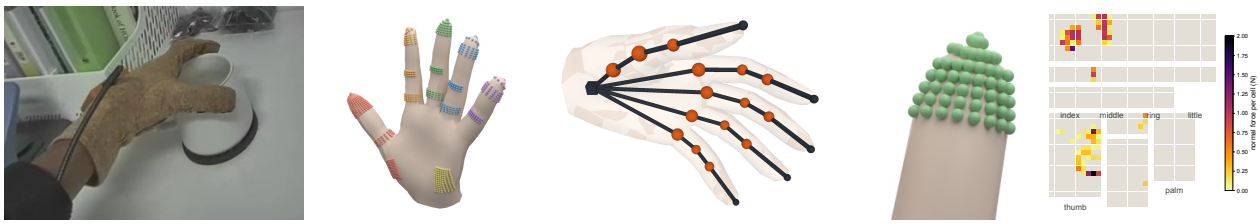
1 Introduction

Human demonstrations are the easiest dexterous manipulation data to collect at scale, and real-to-sim-to-real pipelines rebuild them in simulation, where one recording seeds many (Wang et al., 2026; Mao et al., 2026; Choi et al., 2026; Dan et al., 2025; Allshire et al., 2025). However, these pipelines rebuild the interaction from vision, which recovers only how the hand moves, how the object moves, or both. Contact force never enters

the simulation, since video cannot observe it (Zhou et al., 2026b). A hand pose shows where each finger stops, not how far it closes, which sets grip force, and contacts are often hidden. Video thus misses which fingers touch, when, and how hard, and its two standard uses fail in opposite directions. In our experiments, pose imitation grips with 0.31× the recorded force, and replays built from vision with 1.55×. Physics alone does not recover the human grip either. In our touch-grounded replays of the mug pick-up task, the recorded grip is 1.6–3.7× the smallest force required to hold the mug. Touch measures this quantity (Sundaram et al., 2019; Yang et al., 2022; Song et al., 2025).

Gloves and other wearables have been used to model hand-object dynamics (Zhang et al., 2021), map touch to and from video (Zeng et al., 2026a; Zhang et al., 2026b; Zeng et al., 2026b), supervise robot policies (Luo et al., 2026; Ni et al., 2026; Yu et al., 2024; Xu et al., 2025) and identify object dynamics (Engelbracht et al., 2026), but none rebuilds the hand-object interaction as a physics simulation driven by measured touch, which lets one recording be replayed under new conditions. Touch predicted from video (Zhou et al., 2026b; Zhang et al., 2026a; Zhou et al., 2026a) is trained to reproduce glove signals from the same pixels, and it cannot recover a force the pixels do not show or change the conditions of a recording. In contrast, TouchTwin rebuilds each selected TouchScale (Li et al., 2026) recording in MuJoCo (Todorov et al., 2012), as shown in Figure 1, enabling scalable human-touch data generation. Coding agents write objects as physically tested programs (Fig. 3), a hand model per subject is tracked from both cameras (Potamias et al., 2024) (Fig. 2), and the recorded touch closes each finger past the visible pose to set where, when, and how hard it presses (Fig. 4). In summary, we make the following contributions:

- We propose **TouchTwin**, a Real2Sim2Real pipeline that rebuilds recordings with physically tested hand-object interactions grounded by touch, and simulation carries force along with motion.
- We build a data engine whose replays seed perturbed demonstrations and realistic video with contact and force labels, scaling scarce touch data without additional capture.
- We conduct simulation experiments showing that touch fixes what vision cannot. Pose imitation grips at 0.31× the recorded force, replays built from vision at 1.55×, and replays calibrated by touch at 0.99×, cutting missed contact frames from 32% to 13%.



(a) Record **(b)** Fit **(c)** Rig **(d)** Taxels **(e)** Check

Figure 2 Simulating a tactile hand from egocentric video and paired tactile recordings. (a) A participant's tactile glove grasping a mug, seen by the head camera. (b) The simulated hand fitted to the participant's episodes, with the glove's taxels on its skin. (c) The joint chain, one dot per hinge. (d) A fingertip pad and its taxels. (e) The real glove's force map over its 880 cells, which the simulated touch is checked against.

2 Related Work

Real-to-sim-to-real from human demonstrations. Recent pipelines turn human videos into simulated demonstrations for dexterous hands and humanoids: they reconstruct the hand and object from monocular or egocentric video, make the interaction physically executable, and use the simulation to train or test robot policies (Wang et al., 2026; Mao et al., 2026; Choi et al., 2026; Dan et al., 2025; Allshire et al., 2025). Data generation systems multiply a few demonstrations by perturbing object poses and replaying segments (Mandlekar et al., 2023; Jiang et al., 2025). All of these take the interaction from vision; TouchTwin adds measured touch as the signal that decides contact and grip force.

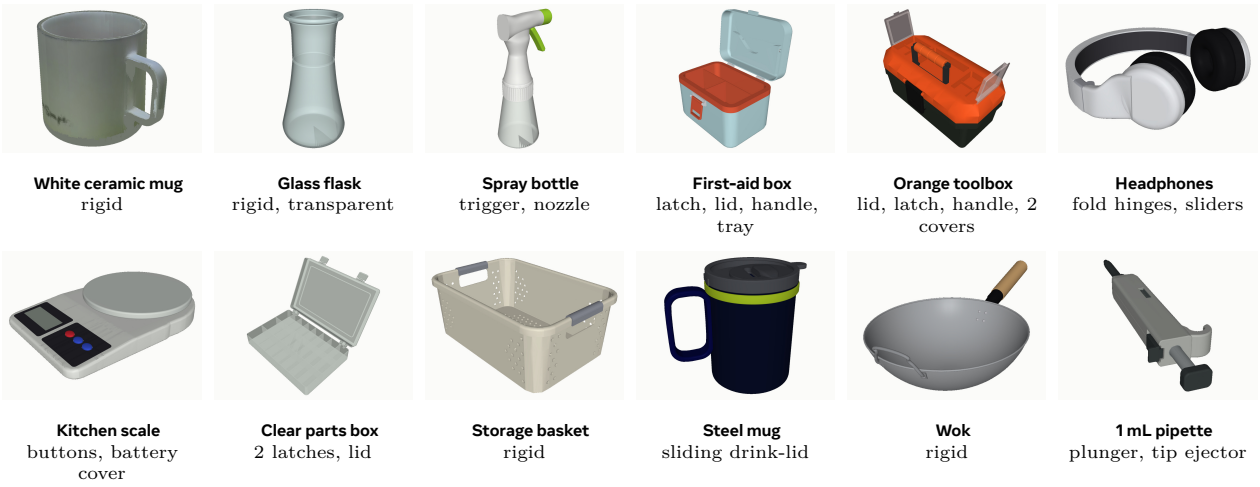


Figure 3 Examples of objects written by coding agents as physics-tested MuJoCo programs. Each tile is a render of one simulation asset; the line under its name lists its moving parts. Interactive versions are on the project page.

Human touch. Tactile gloves record the pressure distribution of the human grasp (Sundaram et al., 2019), and recent datasets pair touch with vision at scale (Yang et al., 2022; Song et al., 2025; Li et al., 2026). Touch has been used to model hand-object dynamics (Zhang et al., 2021), predicted from egocentric video (Zhou et al., 2026b; Zhang et al., 2026a; Zhou et al., 2026a; Zeng et al., 2026a), used to generate hand video (Zhang et al., 2026b; Zeng et al., 2026b), and used to supervise robot policies (Luo et al., 2026; Ni et al., 2026; Yu et al., 2024; Xu et al., 2025; NeoteAI Team and Fudan TEAI Team, 2026) and object digital twins (Engelbracht et al., 2026). These works map between sensor streams of a recording; a simulator grounded by touch can also replay the recording under new conditions.

Simulation-conditioned video generation. Physics engines and code have been used to drive video generation (Lv et al., 2023; Liu et al., 2024; Code Video Model Team, 2026; Chen et al., 2026), and robot-oriented video generation renders simulation as realistic video (Huang et al., 2026). We condition Wan2.2 (Team Wan et al., 2025) on touch-grounded replays of human hands.

3 Experiments

Simulation makes generated video follow the physics. Conditioned on the simulated hand and object, as in code- and physics-driven video generation (Lv et al., 2023; Liu et al., 2024; Code Video Model Team, 2026; Chen et al., 2026; Huang et al., 2026), Wan2.2 (Team Wan et al., 2025) generation comes closest to the real video (Table 1). Without the simulation, the generated mug takes its own path (Fig. 5).

video generation	PSNR↑	SSIM↑	LPIPS↓	DINO↑
sim render only	7.79	0.309	0.757	0.537
first frame + text (baseline)	15.36	0.663	0.489	0.854
baseline + depth control	18.55	0.759	0.276	0.914
baseline + sim control	**18.89**	**0.766**	**0.258**	**0.928**

Table 1 Generated vs. real video (same first frame, prompt and seed).

Touch-grounding is essential for contact accuracy. A contact predictor trained on vision-based simulations misses 32% of the real contact frames on held-out recordings, whereas training on touch-grounded simulations reduces miss rate to 13% (Table 2).

contact prediction, trained on	contact F1↑	missed contact↓	press dur. err.↓
vision-based replays	0.778	32.2%	1.01 s
touch-grounded replays (ours)	**0.890**	**13.1%**	**0.61 s**

Table 2 Contact prediction on held-out human-ego recordings. Press dur. err.: |predicted − true| press time per finger.

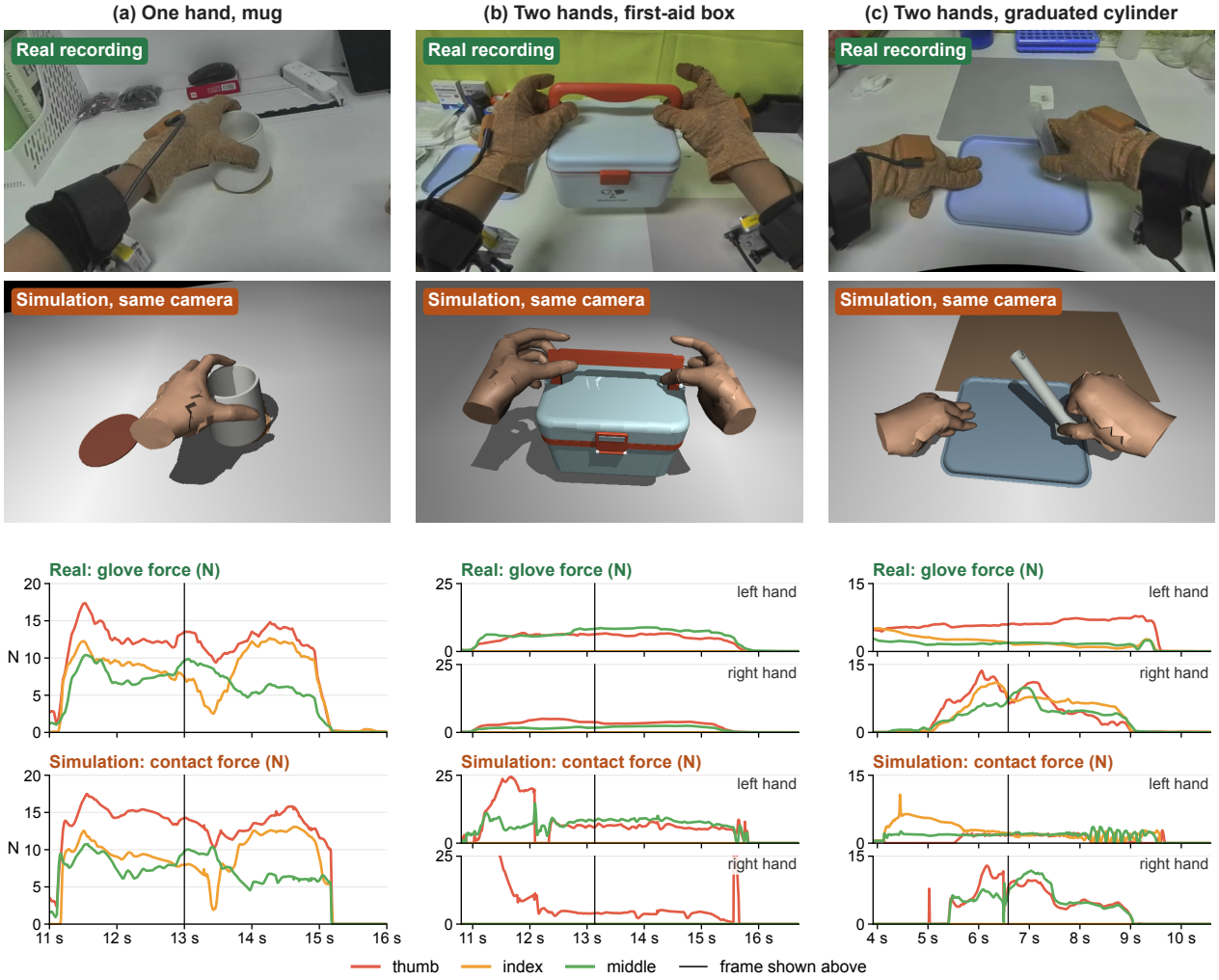


Figure 4 Touch-grounded replays simulate the contact force. One column per clip, real (green) above simulation (orange). Top: a frame of the real recording from the head camera, and the simulation rendered through the same camera. Bottom: the normal force of the thumb, index and middle finger over the clip, measured by the glove and simulated as contact force, on the same axes; the vertical line marks the frame shown. The object is never scripted; it moves only through contact. (a) The left hand moves a mug to a marked spot. (b) Both hands pinch the handle of a first-aid box, lift, carry and set it down; the simulated thumbs overshoot the glove while the grip closes and at release (peaks above 25 N cut off). (c) The right hand lifts a graduated cylinder out of a tray and stands it on a mat while the left hand holds the tray.

Touch supplies information beyond surface-level hand positions. A hand policy (N_0-VTLA (NeoteAI Team and Fudan TEAI Team, 2026)) trained on the observed finger pose grips held-out recordings with 0.31× the true force (Table 3); the touch-grounded replays restore human-level grip (0.99×, 20/20).

hand policy, trained on	held-out recordings↑	grip vs. recording	simulated test↑
observed finger pose	17/20	0.31	26/36
vision-based replays	15/20	1.55	33/36
touch-grounded replays (ours)	**20/20**	**0.99**	**35/36**

Table 3 Hand policy on held-out recordings and simulated tests; success: slip ≤5 mm, set down ≤30 mm off.

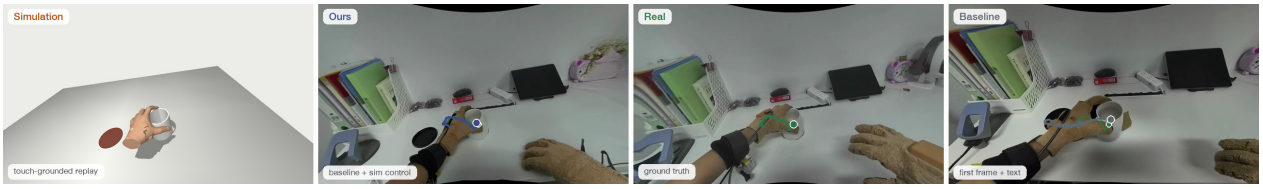


Figure 5 Simulation-conditioned generation. From left: the touch-grounded replay rendered through the recorded head camera, the generated video conditioned on it (baseline + sim control), the real video, and generation from the first frame and text alone, all at the same moment. Same first frame, prompt and seed; lines trace the mug, green = real.

4 Downstream Applications

Each replay also seeds perturbed demonstrations with simulated contact and force labels, rendered as realistic video in new views and scenes.

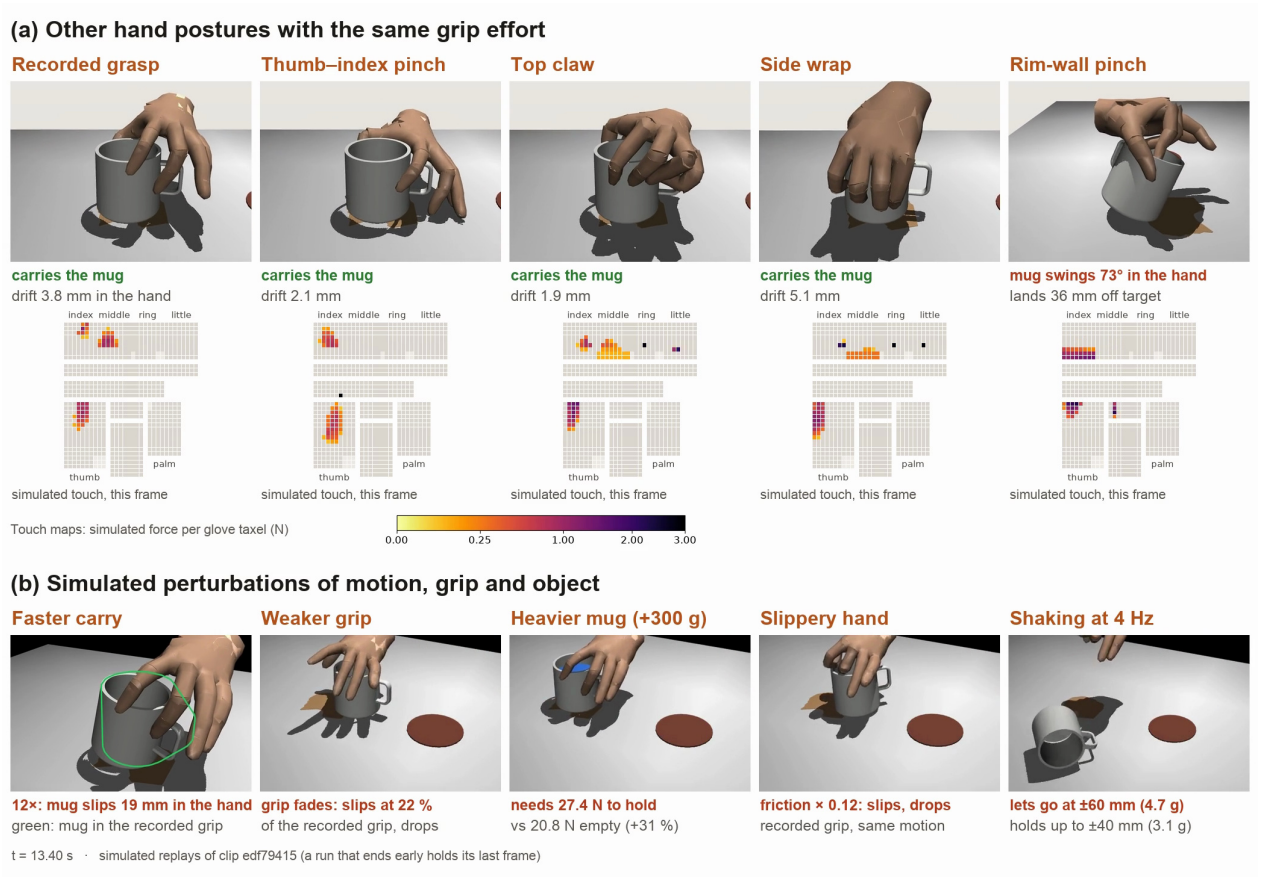
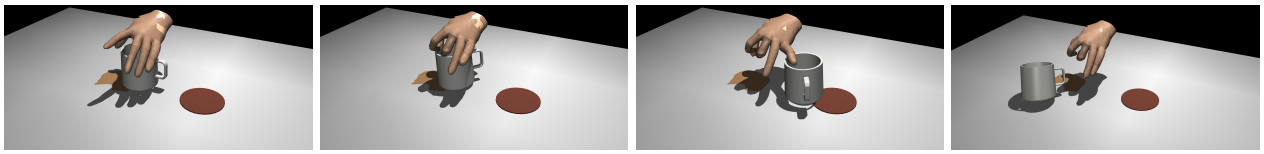


Figure 6 Perturbed replays of one recording. (a) Other hand postures with the same grip effort; under each, the simulated touch of that frame on the glove's pad layout. The recorded grasp, thumb-index pinch, top claw and side wrap carry the mug (drift 3.8, 2.1, 1.9 and 5.1 mm in the hand); with the rim-wall pinch the mug swings 73° in the hand and lands 36 mm off target. (b) Perturbations of motion, grip and object. A 12× faster carry slips the mug 19 mm in the hand (the green outline is the mug in the recorded grip); the grip fades until the mug slips at 22% of the recorded grip and drops; with 300 g of water the mug needs 27.4 N to hold against 20.8 N empty (+31%); at 0.12× friction it slips and drops; a 4 Hz shake makes the hand let go at ±60 mm (4.7 g).

Physics turns one recording into labeled successes and failures. A perturbation changes one thing in a replay and re-runs the physics: the grasp (another hand posture at the same grip effort), the motion (a faster carry or an added shake), the grip (each finger's force target scaled down) or the object (heavier,

or more slippery). With nothing changed, it reproduces the replay exactly. Because physics resolves the outcome, the mug can slip, swing or drop, and every perturbed replay carries simulated per-taxel touch, contact and force in the glove's format (Fig. 6). Sweeping one factor until the object drops measures how much margin the recorded grasp had (Fig. 7).



(a) Grip fades **(b)** Hand gets slippery **(c)** Mug gets heavier **(d)** Shake builds up

Figure 7 Perturbed until the mug drops. During the carry, (a) every finger's force target is turned down from 1× to 0, (b) the hand-mug friction is turned down, (c) mass is added up to 7.5 kg (far more than water could add), and (d) a 4 Hz side-to-side shake grows in amplitude. Each sweep runs until the mug leaves the hand.

One replay becomes realistic video in new views and scenes. Wan2.2 (Team Wan et al., 2025) renders a replay or perturbed replay as realistic video, conditioned on the simulation along the recorded head-camera trajectory. Changing the camera, the hand posture or the scene gives the same simulated interaction from other viewpoints, with other grasps and in other places, and each generated video inherits the contact, touch and force labels of its replay (Fig. 8).



Figure 8 Generated videos decoded from touch-grounded replays (one frame each). From left: a generated video in the recording's own view and scene; a wrist camera in the same scene; new viewpoints, hand postures and scenes.

Simulation measures touch where the glove has no sensors. The glove does not cover the whole hand: the centre of the palm and parts of the finger bases have no sensors. The simulated hand has skin everywhere, so a replay records contact that the glove cannot. Figure 9 perturbs the mug replay so that the palm presses onto the mug: the glove-layout palm taxels stay at zero, while the simulator reports the normal force on palm skin without taxels.

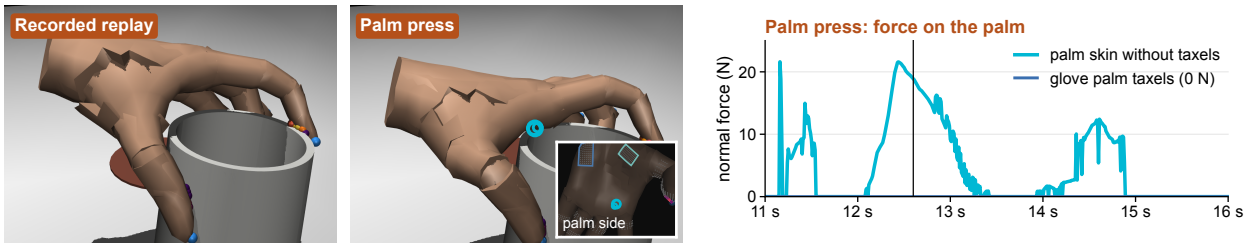


Figure 9 Touch beyond the glove. Left: the recorded replay, where the palm does not touch the mug. Middle: a simulated palm press, with the wrist moved 8 mm along the palm normal past the first visible palm contact and the fingers kept on their spots; inset: the palm side, with the glove's palm pads outlined. Cyan rings: contact on palm skin the glove does not cover. Right: during the press, the palm skin carries up to 22 N while the glove's palm taxels read zero; the vertical line marks the frame shown.

5 Limitations

All policies act on simulated assets and are tested open-loop in simulation; transfer to robot hands, real-robot experiments and task diversification are still needed. The held-out sets are small, the policy and

contact experiments use the mug task only, and the held-out clips start with the hand already in contact, so contact onset is not tested. Two-hand replays still slip 12–23 mm in the hand. Generated videos inherit the simulation's errors, and the video model can still change details the simulation does not control.

Acknowledgments

This work used the VISION supercomputer provided by Texas A&M High Performance Research Computing (HPRC). We thank AMD for providing access to AMD Instinct MI300X GPUs through the AMD University Program and the AMD Developer Cloud.

References

A. Allshire, H. Choi, J. Zhang, D. McAllister, et al. Visual imitation enables contextual humanoid control. In *Conference on Robot Learning (CoRL)*, 2025.

Yiwen Chen, Guosheng Lin, and Chi Zhang. Code world model: Coding agent as world brain. *arXiv preprint arXiv:2608.25927*, 2026.

Byeonggyeol Choi, Woojin Oh, and Jongwoo Lim. Grasp-and-lift: Executable 3D hand-object interaction reconstruction via physics-in-the-loop optimization. *arXiv preprint arXiv:2601.18121*, 2026.

Code Video Model Team. Code Video Model: Conditioning video generation on code. Technical report, https://code-video-model.github.io/tech_report.pdf, 2026.

P. Dan, K. Kedia, et al. X-Sim: Cross-embodiment learning via real-to-sim-to-real. In *Conference on Robot Learning (CoRL)*, 2025.

T. Engelbracht et al. ForceTwin: Physics-informed digital twins for robotic manipulation from instrumented human interaction. *arXiv preprint arXiv:2609.21751*, 2026.

Huang Huang, Wensi Ai, Ziyu Chen, Youhui Wang, et al. RoboRender: Robot-oriented video generation for visual sim-to-real transfer. *arXiv preprint arXiv:2610.09254*, 2026.

Zhenyu Jiang, Yuqi Xie, Kevin Lin, Zhenjia Xu, Weikang Wan, Ajay Mandlekar, Linxi Fan, and Yuke Zhu. DexMimicGen: Automated data generation for bimanual dexterous manipulation via imitation learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

Dayou Li, Hao Wang, Qianqian Yang, Zihao Zhu, et al. TouchScale: 500 hours of human vision and touch for visual-tactile learning. *arXiv preprint arXiv:2610.10288*, 2026.

Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. PhysGen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision (ECCV)*, 2024.

Shengcheng Luo, Xiaoyang Cheng, Hong Ying, et al. Touch2Robot: Robot touch in the human demonstration loop. *arXiv preprint arXiv:2609.24660*, 2026.

Jiaxi Lv, Yi Huang, Mingfu Yan, Jiancheng Huang, et al. GPT4Motion: Scripting physical motions in text-to-video generation via Blender-oriented GPT planning. *arXiv preprint arXiv:2311.12631*, 2023.

Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. MimicGen: A data generation system for scalable robot learning using human demonstrations. In *Conference on Robot Learning (CoRL)*, 2023.

Ting Mao, Yanming Shao, Ziheng Wang, et al. OmniHOI: Dexterous hand-object interaction from monocular human video. *arXiv preprint arXiv:2610.10855*, 2026.

NeoteAI Team and Fudan TEAI Team. N_0-VTLA: Scaling vision-tactile-language-action model with latent tactile tokens. *arXiv preprint arXiv:2607.23782*, 2026.

S. Ni, H. Zhang, Z. Wei, et al. TactiDex: A real-world tactile-guided benchmark for human-like dexterous manipulation. *arXiv preprint arXiv:2607.09190*, 2026.

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, et al. WiLoR: End-to-end 3D hand localization and reconstruction in-the-wild. *arXiv preprint arXiv:2409.12259*, 2024.

Yuxin Ray Song, Jinzhou Li, Rao Fu, Devin Murphy, et al. OpenTouch: Bringing full-hand touch to real-world interaction. *arXiv preprint arXiv:2512.16842*, 2025.

Subramanian Sundaram, Petr Kellnhofer, Yunzhu Li, Jun-Yan Zhu, Antonio Torralba, and Wojciech Matusik. Learning the signatures of the human grasp using a scalable tactile glove. *Nature*, 569:698–702, 2019.

Team Wan, Ang Wang, Baole Ai, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5026–5033, 2012.

Youhui Wang, Yunzhu Li, Li Fei-Fei, et al. DexAgent: An agentic Human2Sim2Robot framework for dexterous manipulation with self-evolving tool library. *arXiv preprint arXiv:2609.35318*, 2026.

M. Xu, H. Zhang, et al. DexUMI: Using human hand as the universal manipulation interface for dexterous manipulation. In *Conference on Robot Learning (CoRL)*, 2025.

Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. Touch and Go: Learning from human-collected vision and touch. In *NeurIPS Datasets and Benchmarks Track*, 2022.

K. Yu, Y. Han, Q. Wang, V. Saxena, D. Xu, and Y. Zhao. MimicTouch: Leveraging multi-modal human tactile demonstrations for contact-rich manipulation. In *Conference on Robot Learning (CoRL)*, 2024.

Yuan Zeng, Yujia Shi, Tiao Tan, et al. EgoTactile: Learning grasp pressure for everyday objects from egocentric video. In *International Conference on Machine Learning (ICML)*, 2026a.

Ziyao Zeng, Xiatao Sun, Hao Wang, et al. Dexterous tactile world model. *arXiv preprint arXiv:2609.34286*, 2026b.

Qiang Zhang, Yunzhu Li, Yiyue Luo, Wan Shou, Michael Foshey, Junchi Yan, Joshua B. Tenenbaum, Wojciech Matusik, and Antonio Torralba. Dynamic modeling of hand-object interactions via tactile sensing. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021.

Wenkang Zhang, Chengbo Yuan, Zicheng Zhang, et al. EgoTac: In-the-wild tactile prediction from egocentric vision. *arXiv preprint arXiv:2608.15060*, 2026a.

Xinyu Zhang, Ziyi Kou, Chuan Qin, et al. Glove2Hand: Synthesizing natural hand-object interaction from multi-modal sensing gloves. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026b.

Danyan Zhou, Jinxuan Lu, Jiawei Lin, et al. TouchSight: Bare-handed tactile prediction from egocentric video via generative visual augmentation. *arXiv preprint arXiv:2609.20414*, 2026a.

Jianyi Zhou et al. TouchAnything: A dataset and framework for bimanual tactile estimation from egocentric video. *arXiv preprint arXiv:2605.13083*, 2026b.