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

Abstract: 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: a hand policy that imitates the observed pose grips held-out objects 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. Project page with videos: https://touchtwin.github.io.

Keywords: Real-to-sim-to-real, Dexterous manipulation, Egocentric data

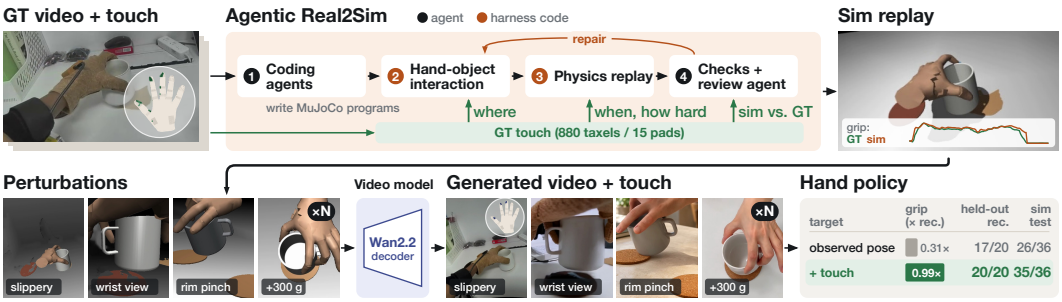


Figure 1: Top: an agentic pipeline rebuilds each recording; GT touch sets where, when and how hard fingers touch and checks the replay. Bottom: each replay yields N perturbed demos, decoded to video in new views and scenes, that train a hand policy.

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 [1–5]. However, these pipelines rebuild the interaction from vision, which recovers only how

10th Conference on Robot Learning (CoRL 2026), Austin TX, USA.

the hand moves, how the object moves, or both. Contact force never enters the simulation, since video cannot observe it [6]. 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 [7–9].

Gloves and other wearables have been used to model hand-object dynamics [10], map touch to and from video [11–13], supervise robot policies [14–17] and identify object dynamics [18], 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 [6, 19, 20] 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 [21] recording in MuJoCo [22], as shown in Figure 1, enabling scalable human-touch data generation. Coding agents write objects as physically tested programs, a hand model per subject is tracked from both cameras [23], and the recorded touch closes each finger past the visible pose to set where, when, and how hard it presses. 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%.

2 Results

Simulation makes generated video follow the physics. Conditioned on the simulated hand and object, as in code- and physics-driven video generation [24–28], Wan2.2 [29] generation comes closest to the real video (Table 1).

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.

Touch supplies information beyond surface-level hand positions. A hand policy (N_0-VTLA [30]) 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.

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.

References

[1] Y. Wang, Y. Li, L. Fei-Fei, et al. DexAgent: An agentic Human2Sim2Robot framework for dexterous manipulation with self-evolving tool library. *arXiv preprint arXiv:2609.35318*, 2026.

[2] T. Mao, Y. Shao, Z. Wang, et al. OmniHOI: Dexterous hand-object interaction from monocular human video. *arXiv preprint arXiv:2610.10855*, 2026.

[3] B. Choi, W. Oh, and J. Lim. Grasp-and-lift: Executable 3D hand-object interaction reconstruction via physics-in-the-loop optimization. *arXiv preprint arXiv:2601.18121*, 2026.

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

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

[6] J. Zhou et al. TouchAnything: A dataset and framework for bimanual tactile estimation from egocentric video. *arXiv preprint arXiv:2605.13083*, 2026.

[7] S. Sundaram, P. Kellnhofer, Y. Li, J.-Y. Zhu, A. Torralba, and W. Matusik. Learning the signatures of the human grasp using a scalable tactile glove. *Nature*, 569:698–702, 2019.

[8] F. Yang, C. Ma, J. Zhang, J. Zhu, W. Yuan, and A. Owens. Touch and Go: Learning from human-collected vision and touch. In *NeurIPS Datasets and Benchmarks Track*, 2022.

[9] Y. R. Song, J. Li, R. Fu, D. Murphy, et al. OpenTouch: Bringing full-hand touch to real-world interaction. *arXiv preprint arXiv:2512.16842*, 2025.

[10] Q. Zhang, Y. Li, Y. Luo, W. Shou, M. Foshey, J. Yan, J. B. Tenenbaum, W. Matusik, and A. Torralba. Dynamic modeling of hand-object interactions via tactile sensing. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021.

[11] Y. Zeng, Y. Shi, T. Tan, et al. EgoTactile: Learning grasp pressure for everyday objects from egocentric video. In *International Conference on Machine Learning (ICML)*, 2026.

[12] X. Zhang, Z. Kou, C. 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)*, 2026.

[13] Z. Zeng, X. Sun, H. Wang, et al. Dexterous tactile world model. *arXiv preprint arXiv:2609.34286*, 2026.

[14] S. Luo, X. Cheng, H. Ying, et al. Touch2Robot: Robot touch in the human demonstration loop. *arXiv preprint arXiv:2609.24660*, 2026.

[15] 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.

[16] 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.

[17] 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.

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

[19] W. Zhang, C. Yuan, Z. Zhang, et al. EgoTac: In-the-wild tactile prediction from egocentric vision. *arXiv preprint arXiv:2608.15060*, 2026.

[20] D. Zhou, J. Lu, J. Lin, et al. TouchSight: Bare-handed tactile prediction from egocentric video via generative visual augmentation. *arXiv preprint arXiv:2609.20414*, 2026.

[21] D. Li, H. Wang, Q. Yang, Z. Zhu, et al. TouchScale: 500 hours of human vision and touch for visual-tactile learning. *arXiv preprint arXiv:2610.10288*, 2026.

[22] E. Todorov, T. Erez, and Y. Tassa. MuJoCo: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5026–5033, 2012.

[23] R. A. Potamias, J. Zhang, J. Deng, et al. WiLoR: End-to-end 3D hand localization and reconstruction in-the-wild. *arXiv preprint arXiv:2409.12259*, 2024.

[24] J. Lv, Y. Huang, M. Yan, J. Huang, et al. GPT4Motion: Scripting physical motions in text-to-video generation via Blender-oriented GPT planning. *arXiv preprint arXiv:2311.12631*, 2023.

[25] S. Liu, Z. Ren, S. Gupta, and S. Wang. PhysGen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision (ECCV)*, 2024.

[26] Z. Liang, J. Zhang, G. Qian, D. Chen, and Z. Wan. Code video model, Oct. 2026. URL https://code-video-model.github.io/tech_report.pdf. Preprint.

[27] Y. Chen, G. Lin, and C. Zhang. Code world model: Coding agent as world brain. *arXiv preprint arXiv:2608.25927*, 2026.

[28] H. Huang, W. Ai, Z. Chen, Y. Wang, et al. RoboRender: Robot-oriented video generation for visual sim-to-real transfer. *arXiv preprint arXiv:2610.09254*, 2026.

[29] Team Wan, A. Wang, B. Ai, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[30] 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.