UniVTAC: A Unified Simulation Platform for Visuo-Tactile Manipulation Data Generation, Learning, and Benchmarking

Baijun Chen[5,2*], Weijie Wan[6*], Tianxing Chen[4*], Xianda Guo[7,2*], Congsheng Xu[1], Yuanyang Qi[3], Haojie Zhang[3], Longyan Wu[8], Tianling Xu[1], Zixuan Li[6], Yizhe Wu[3], Rui Li[3], Xiaokang Yang[1], Ping Luo[4], Wei Sui[2,†], and Yao Mu[1,†]

[1] ScaleLab, Shanghai Jiao Tong University [2] D-Robotics [3] ViTai Robotics [4] The University of Hong Kong [5] Nanjing University [6] Shenzhen University [7] Wuhan University [8] Fudan University
*Equal Contribution †Corresponding Authors

***Abstract*—Visuo-tactile perception is crucial for contact-rich manipulation, yet scalable tactile data collection and standardized policy evaluation remain difficult due to hardware cost, sensor heterogeneity, and limited benchmarks. We present UniVTAC, a simulation-based framework for visuo-tactile data generation, representation learning, and benchmarking. UniVTAC supports multiple optical tactile sensors, tactile-aware manipulation primitives, and privileged physical annotations such as marker-free tactile images, depth maps, marker displacements, and object poses. It further provides eight contact-rich manipulation tasks spanning shape perception, pose reasoning, and contact-rich interaction. As a proof of utility, we pretrain a visuo-tactile encoder from simulation-generated data and integrate it into policy learning. Experiments show that UniVTAC improves average success from 30.9% to 48.0% in simulation and from 43.3% to 68.3% in real-world tasks.**

***Index Terms*—visuo-tactile perception, simulation, benchmark, robotic manipulation**

I. Introduction

Visuo-tactile perception is critical for contact-rich manipulation, where visual observations are often limited by occlusion, close-range depth noise, and the lack of direct contact-state feedback. Tactile sensing provides complementary information about local geometry, deformation, and relative motion, enabling policies to detect misalignment and perform corrective actions during insertion, alignment, and grasping.

Despite recent progress in tactile sensor simulation [1]–[8] and simulation-based robot learning [9]–[17], scalable infrastructure for visuo-tactile manipulation remains underdeveloped. Real-world tactile data collection is costly and difficult to standardize across sensors, while existing manipulation benchmarks provide limited support for tactile-rich contact dynamics and unified evaluation. This limits both tactile representation learning and fair comparison of tactile-driven policies.

We propose UniVTAC, a unified simulation framework for visuo-tactile data generation, representation learning, and benchmarking. UniVTAC supports three optical tactile sensors, provides tactile-aware manipulation primitives, and generates privileged physical annotations including marker-free tactile images, depth maps, marker displacements, and object poses. Based on these signals, we pretrain a visuo-tactile encoder and evaluate it on an eight-task contact-rich benchmark. Experiments show that the pretrained encoder improves the average simulation success rate from 30.9% to 48.0%, and improves real-world success from 43.3% to 68.3%.

Our contributions are: (1) a unified visuo-tactile simulation framework for scalable tactile data generation; (2) a privileged-supervision pipeline for tactile-centric representation learning; and (3) an eight-task contact-rich simulation benchmark that enables systematic and reproducible evaluation of visuo-tactile manipulation policies.

II. UniVTAC

UniVTAC is an end-to-end simulation framework for visuo-tactile manipulation, covering sensor simulation, tactile-aware data generation, representation learning, and policy evaluation.

Platform and data generation. Built upon TacEx [4] and Isaac Sim [20], UniVTAC extends soft-body visuo-tactile simulation to three visuo-tactile sensors: *GelSight Mini* [21], *ViTai GF225* [22], and *Xense WS* [23]. The sensor geometry, camera intrinsics, gelpad mesh, and rendering configuration are modeled in a modular manner, allowing the same manipulation pipeline to be evaluated across different tactile hardware. To enable scalable data collection, UniVTAC provides atomic manipulation primitives including *Grasp*, *Move*, *Place*, *Probe*, and *Rotate*. In particular, *Grasp* and *Probe* use tactile-reactive feedback to avoid unrealistic penetration and produce physically meaningful contact observations.

Privileged supervision and encoder pretraining. A key advantage of simulation is access to physical signals that are difficult to obtain in the real world. UniVTAC records raw marker-based tactile images together with marker-free tactile images, gelpad depth maps, marker displacements, and object poses in the gelpad frame. These privileged annotations supervise three complementary perception pathways: shape reconstruction, contact deformation prediction, and pose

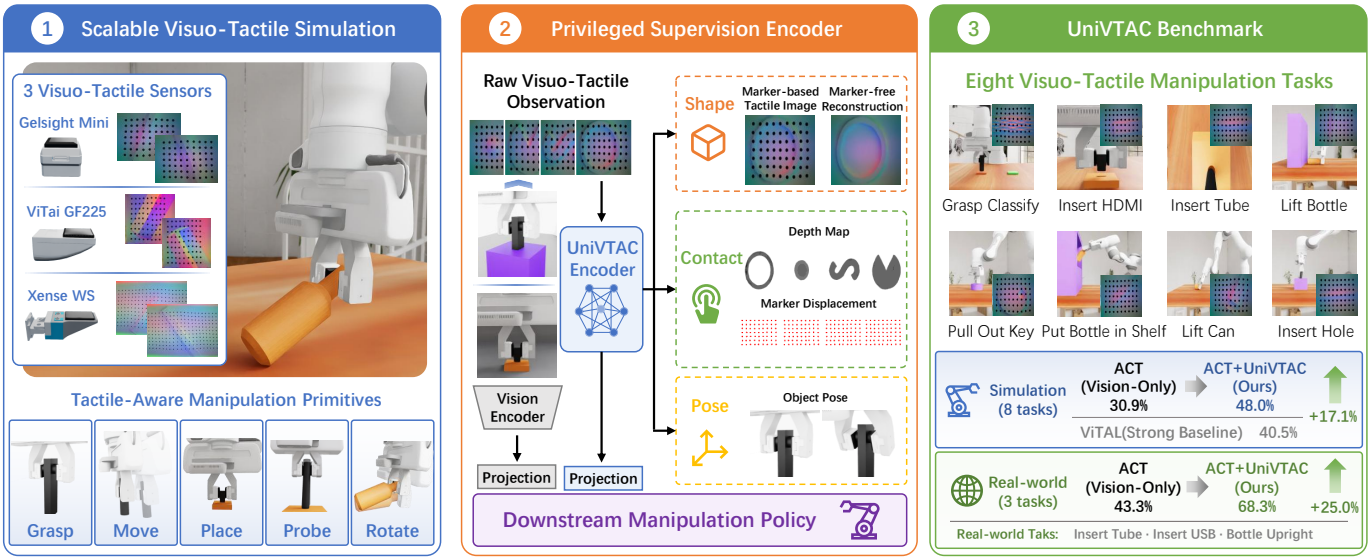

Fig. 1. Overview of our UniVTAC platform.

TABLE I

UNIVTAC BENCHMARK. WE REPORT THE SUCCESS RATES AND AVERAGE PERFORMANCE OF ACT WITHOUT TACTILE INPUT, VITAL, AND ACT WITH THE UNIVTAC ENCODER (OURS) ACROSS THE EIGHT TASKS IN THE UNIVTAC BENCHMARK. (**BEST**, SECOND-BEST)

Method	Lift Bottle	Pull-out Key	Lift Can	Put Bottle in Shelf	Insert Hole	Insert HDMI	Insert Tube	Grasp Classify	*Average*
ACT [18]	42.0	28.0	20.0	28.0	19.0	15.0	45.0	50.0	30.9
VITaL [19]	**72.0**	**47.0**	8.0	**32.0**	**25.0**	6.0	34.0	**100.0**	40.5
Ours	71.0	46.0	**29.0**	31.0	24.0	**28.0**	**56.0**	99.0	**48.0**

regression. As shown in Fig. 1, the UniVTAC Encoder is pretrained with these auxiliary objectives, while the decoder heads are discarded during policy learning, introducing no additional inference-time overhead.

Contact-Rich Manipulation Benchmark. UniVTAC further provides an eight-task benchmark for evaluating tactile-dependent manipulation policies. As shown in Fig. 1, the tasks are organized into three categories: **shape perception**, where *Grasp Classify* requires inferring object geometry from tactile observations; **pose reasoning**, where *Lift Bottle*, *Lift Can*, and *Put Bottle in Shelf* require estimating object pose and relative spatial relationships; and **contact-rich interaction**, where *Insert Hole*, *Insert Tube*, *Insert HDMI*, and *Pull Out Key* require fine-grained alignment and contact-based correction. To make tactile observations informative, expert trajectories include pose randomization and corrective behaviors, producing diverse contact patterns rather than near-perfect collision-free motions.

III. EXPERIMENTS

We evaluate whether UniVTAC-generated data and privileged supervision improve downstream manipulation. In simulation, we compare ACT [18], VITaL [19], and ACT augmented with the UniVTAC Encoder on the eight-task UniVTAC Benchmark. All policies are trained with 50 automatically collected demonstrations per task and evaluated over 100 rollouts. As shown in Table I, UniVTAC improves the average success rate from 30.9% for vision-only ACT to 48.0%, and also outperforms the visuo-tactile pretraining baseline VITaL, which achieves 40.5%.

We further evaluate the sim-to-real transfer of the UniVTAC Encoder on three real-world tasks using the ViTai GF225 visuo-tactile sensor: *Insert Tube*, *Insert USB*, and *Bottle Upright*. The UniVTAC Encoder is first pretrained purely in simulation with privileged supervision, and is then integrated into ACT and trained on real-world demonstrations together with the policy. The results show that ACT with the UniVTAC Encoder improves the average real-world success rate from 43.3% to 68.3% compared with the vision-only ACT baseline, demonstrating the effectiveness of the simulation-pretrained tactile encoder for real-world contact-rich manipulation.

IV. CONCLUSION

We introduced UniVTAC, a unified simulation framework for scalable visuo-tactile data generation, representation learning, and benchmarking. By combining visuo-tactile sensor simulation, privileged physical supervision, and eight contact-rich manipulation tasks, UniVTAC enables systematic evaluation of tactile-driven policies. Experiments in simulation and on real robots show that UniVTAC-pretrained representations improve manipulation success, highlighting the potential of simulation-based tactile data for contact-rich robot learning.

REFERENCES

[1] M. Li, Z. Ferguson, T. Schneider, T. Langlois, D. Zorin, D. Panozzo, C. Jiang, and D. M. Kaufman, "Incremental potential contact: intersection-and inversion-free, large-deformation dynamics," *ACM transactions on graphics*, 2020.

[2] I. Akinola, J. Xu, J. Carius, D. Fox, and Y. Narang, "TacSL: A Library for Visuotactile Sensor Simulation and Learning," *IEEE Transactions on Robotics*, vol. 41, pp. 2645–2661, 2025.

[3] C. Zhang, S. Cui, J. Hu, T. Jiang, T. Zhang, R. Wang, and S. Wang, "TacFlex: Multimode Tactile Imprints Simulation for Visuotactile Sensors With Coating Patterns," *IEEE Transactions on Robotics*, vol. 41, pp. 3965–3985, 2025.

[4] D. H. Nguyen, T. Schneider, G. Duret, A. Kshirsagar, B. Belousov, and J. Peters, "TacEx: GelSight Tactile Simulation in Isaac Sim – Combining Soft-Body and Visuotactile Simulators," 2024-11-07.

[5] Z. Si, G. Zhang, Q. Ben, B. Romero, Z. Xian, C. Liu, and C. Gan, "Difftactile: A physics-based differentiable tactile simulator for contact-rich robotic manipulation," *arXiv preprint arXiv:2403.08716*, 2024.

[6] Z. Chen, S. Zhang, S. Luo, F. Sun, and B. Fang, "Tacchi: A Pluggable and Low Computational Cost Elastomer Deformation Simulator for Optical Tactile Sensors," *IEEE Robotics and Automation Letters*, vol. 8, no. 3, pp. 1239–1246, 2023-03.

[7] Y. Li, W. Du, C. Yu, P. Li, Z. Zhao, T. Liu, C. Jiang, Y. Zhu, and S. Huang, "Taccel: Scaling Up Vision-based Tactile Robotics via High-performance GPU Simulation," 2025-04-17.

[8] W. Chen, J. Xu, F. Xiang, X. Yuan, H. Su, and R. Chen, "General-Purpose Sim2Real Protocol for Learning Contact-Rich Manipulation With Marker-Based Visuotactile Sensors," *IEEE Transactions on Robotics*, vol. 40, pp. 1509–1526, 2024.

[9] F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang, *et al.*, "Sapien: A simulated part-based interactive environment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11097–11107, 2020.

[10] J. Gu, F. Xiang, X. Li, Z. Ling, X. Liu, T. Mu, Y. Tang, S. Tao, X. Wei, Y. Yao, *et al.*, "Maniskill2: A unified benchmark for generalizable manipulation skills," in *The Eleventh International Conference on Learning Representations*, 2023.

[11] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine, "Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning," in *Conference on robot learning*, pp. 1094–1100, PMLR, 2020.

[12] O. Mees, L. Hermann, E. Rosete-Beas, and W. Burgard, "Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 7327–7334, 2022.

[13] S. Nasiriany, A. Maddukuri, L. Zhang, A. Parikh, A. Lo, A. Joshi, A. Mandlekar, and Y. Zhu, "Robocasa: Large-scale simulation of everyday tasks for generalist robots," in *Robotics: Science and Systems (RSS)*, 2024.

[14] H. Geng, F. Wang, S. Wei, Y. Li, B. Wang, B. An, C. T. Cheng, H. Lou, P. Li, Y.-J. Wang, *et al.*, "Roboverse: Towards a unified platform, dataset and benchmark for scalable and generalizable robot learning," *arXiv preprint arXiv:2504.18904*, 2025.

[15] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone, "Libero: Benchmarking knowledge transfer for lifelong robot learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 44776–44791, 2023.

[16] T. Chen, Z. Chen, B. Chen, Z. Cai, Y. Liu, Z. Li, Q. Liang, X. Lin, Y. Ge, Z. Gu, *et al.*, "Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation," *arXiv preprint arXiv:2506.18088*, 2025.

[17] Y. Mu, T. Chen, Z. Chen, S. Peng, Z. Lan, Z. Gao, Z. Liang, Q. Yu, Y. Zou, M. Xu, *et al.*, "Robotwin: Dual-arm robot benchmark with generative digital twins," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 27649–27660, 2025.

[18] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware," Apr. 2023.

[19] A. George, S. Gano, P. Katragadda, and A. B. Farimani, "VITaL Pretraining: Visuo-Tactile Pretraining for Tactile and Non-Tactile Manipulation Policies," 2024-09-26.

[20] NVIDIA, "Isaac Sim," 2025.

[21] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, p. 2762, 2017.

[22] V. Robotics, "Vitai-sdk-release." https://github.com/ViTai-Tech/ViTai-SDK-Release, 2026.

[23] "Xense." https://www.xenserobotics.com/product/367/detail/9, 2025. 2.