Beyond Reconstruction: A Real2Sim2Real Harness for Robot Policy Improvement

Parv Kapoor Zifu Wan Kaylee Kirkelie Srujan Deolasee Ashish Kapoor Sai Vemprala
General Robotics

Abstract: Can foundation-model-generated simulations improve real-world robot policies? We present an end-to-end Real2Sim2Real framework combining agentic scene reconstruction, interaction-aware asset completion, and simulation-based policy adaptation. Our harness orchestrates pretrained perception, reasoning, and 3D generation models with geometric and physical validation to construct diverse, interactive manipulation environments. We use frozen robot policies as behavioral discriminators to identify and correct real–sim discrepancies before training residual controllers in simulation. Evaluated across two independently fine-tuned $\pi_{0.5}$ policies, distinct data-collection paradigms, and manipulation tasks, our approach improves simulated camera-pickup success from 2/24 to 14/24 and plug-insertion success from 7/20 to 13/20 in simulation and from 6/20 to 11/20 on hardware. These preliminary results demonstrate the potential of validated, policy-aligned simulations for improving real-world manipulation. Project website: https://genrobo.github.io/r2s2r/.

Keywords: Real2Sim2Real, foundation models, policy evaluation

1 Introduction

Foundation models have made image-to-3D reconstruction increasingly accessible [1], yet turning these reconstructions into useful robot learning environments remains challenging. Static model pipelines propagate errors in segmentation, geometry, and object placement without validating their physical consequences, while generated assets often lack the articulation and contact mechanics required for manipulation. More fundamentally, even a geometrically plausible simulator may elicit different behavior from a pretrained robot policy due to mismatches in camera observations, robot embodiment, or state representation. We therefore use agreement between a frozen policy's actions under corresponding real and simulated observations as a task-specific criterion for simulator calibration [2]. We combine interaction-aware reconstruction with frozen-policy-guided simulator alignment, then train residual controllers in the resulting environments and evaluate their transfer to two real-world manipulation policies, closing the Real2Sim2Real loop [3].

2 Approach

1. Geometric and physical validity: Agentic real-to-sim reconstruction. Building on SimFoundry [4], we develop an agent-driven reconstruction harness that integrates Depth Anything V3 [5] for metric depth, SAM3 [6] for segmentation, Hunyuan3D 2.1 [7] for geometry and appearance, FoundationPose [8] for object registration, and GPT-6 Astra [9] for visual reasoning and occlusion completion. Unlike static model pipelines, Claude orchestrates twelve independently checkpointed stages through an MCP [10, 11] server, inspecting intermediate outputs and enforcing five review gates with operator input. Automated checks validate object completeness, image registra-

tion, scale consistency, collision penetration, and physical settling, enabling early error detection and recovery from verified checkpoints.

2. Interaction validity: From reconstructed meshes to functional mechanisms. Since image-based reconstruction does not reliably recover hidden articulation or contact geometry, we introduce an agent-driven interaction-completion stage. Given natural-language descriptions of object functionality, the agent generates structured mechanism specifications that fit parametric hinges, socket cavities, or threaded contacts to reconstructed assets while preserving their appearance. Physical properties are initialized from geometric estimates and supplied priors, and completed mechanisms are evaluated through interaction-specific physics probes. These tests establish functional behavior of the authored mechanisms, rather than identification of their real-world physical parameters.

3. Behavioral fidelity: Policy-guided simulator alignment and adaptation. Geometrically and physically valid simulations may still induce incorrect policy behavior due to discrepancies in rendering, camera calibration, embodiment, or state representation. We therefore use the frozen pretrained $\pi_{0.5}$ [12] policy itself as a task-aware discriminator of real–sim fidelity. At corresponding real and simulated states, we compare action predictions under matched proprioception and task instructions, using behavioral discrepancies to guide targeted simulator corrections. Crucially, our objective is not to maximize policy success in simulation, but to reproduce its real-world behavior, including failure modes. After simulator alignment, we train task-specific residual controllers in simulation while keeping each base policy frozen: a small MLP regresses a privileged oracle's bounded end-effector corrections from frozen policy features and robot state, never observing object poses, so all inputs remain available on the real robot.

3 Experimental Results

Reconstruction and interaction validity. We demonstrate our harness across diverse real-world environments, including multi-object tabletop scenes, desktop workspaces, robot manipulation workcells, and contact-rich objects such as power strips, cardboard boxes, and threaded jars. Beyond visual reconstruction, these environments support physical interactions spanning object pickup, articulated box folding, plug insertion, and bimanual jar opening. The harness detects and rejects geometrically or physically inconsistent reconstructions, while interaction-specific probes validate the completed mechanisms.

Policy-guided simulator alignment. Despite passing geometric and physical validation, our initial simulator failed to reproduce the pretrained policy's real-world behavior: the policy initiated manipulation in only half of simulated observations, compared to all real observations. Using the frozen policy as a behavioral discriminator, we systematically identified and corrected discrepancies in camera appearance, viewpoint, robot embodiment, and state feedback. Appearance corrections restored action initiation from 6/12 to 12/12 trials, while geometric calibration eliminated systematic grasp biases of 2.5–4 cm. These corrections enabled the previously unreliable policy to consistently grasp (12/12) and lift (10/12) objects in simulation. Transport and insertion remained unreliable, motivating residual policy adaptation.

Simulation-based residual policy adaptation. We evaluate our Real2Sim2Real framework across two distinct manipulation tasks, data-collection paradigms, and independently fine-tuned $\pi_{0.5}$ checkpoints. For plug insertion, demonstrations are collected using a GELLO-style teleoperation device [13], while camera pickup and placement uses a UMI-style handheld data-collection system [14]. The pickup head reads frozen $\pi_{0.5}$ backbone features and the base action, correcting the end effector by up to 6 cm and 14°; the insertion head adds frozen visual features and joint state, applying a gated lateral fingertip shift of up to 20 mm, trained on 80 oracle episodes. Across 24 unseen object placements, the pickup residual improves grasp-and-hold success from 2/24 to 14/24 in simulation, with real-robot results available on our website.[1] For plug insertion, success increases from 7/20 to 13/20 in simulation and from 6/20 to 11/20 on hardware.

[1] https://genrobo.github.io/r2s2r/

References

[1] R. Luo, Y. Liu, D. Cao, M. Tancik, S. Buch, C. Stearns, and W. Xian. SceneRig: An agentic system for simulation-ready 3D scene reconstruction from single images. *arXiv preprint arXiv:2610.28585*, 2026. URL https://arxiv.org/abs/2610.28585.

[2] X. Li, K. Hsu, J. Gu, K. Pertsch, O. Mees, H. R. Walke, C. Fu, I. Lunawat, I. Sieh, S. Kirmani, S. Levine, J. Wu, C. Finn, H. Su, Q. Vuong, and T. Xiao. Evaluating real-world robot manipulation policies in simulation. In P. Agrawal, O. Kroemer, and W. Burgard, editors, *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pages 3705–3728. PMLR, 2025. URL https://proceedings.mlr.press/v270/li25c.html.

[3] M. Torne Villasevil, A. Simeonov, Z. Li, A. Chan, T. Chen, A. Gupta, and P. Agrawal. Reconciling reality through simulation: A real-to-sim-to-real approach for robust manipulation. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024. doi:10.15607/RSS.2024.XX.015.

[4] N. Ranawaka, J. Wong, W.-L. Pai, W.-T. Chu, T. Dai, M. Moghani, H. Yin, Y. Jiang, W. Durbano, B. Huynh, Y. Fang, D. Xu, R. Zhang, L. Fei-Fei, L. Fan, B. Wen, A. Mandlekar, and Y. Zhu. SimFoundry: Modular and automated scene generation for policy learning and evaluation. *arXiv preprint arXiv:2606.28276*, 2026. URL https://arxiv.org/abs/2606.28276.

[5] H. Lin, S. Chen, J. H. Liew, D. Y. Chen, Z. Li, Y. Zhao, S. Peng, H. Guo, X. Zhou, G. Shi, J. Feng, and B. Kang. Depth anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations (ICLR)*, pages 141261–141285, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e4cd50120b6d7e8daff1749d6bbaa889-Abstract-Conference.html.

[6] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, A. Huang, J. Lei, T. Ma, B. Guo, A. Kalla, M. Marks, J. Greer, M. Wang, P. Sun, R. Rädle, T. Afouras, E. Mavroudi, K. Xu, T.-H. Wu, Y. Zhou, L. Momeni, R. Hazra, S. Ding, S. Vaze, F. Porcher, F. Li, S. Li, A. Kamath, H. K. Cheng, P. Dollár, N. Ravi, K. Saenko, P. Zhang, and C. Feichtenhofer. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations (ICLR)*, pages 138846–138923, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e0982cbc81401df3430ee1ff780dc7a2-Abstract-Conference.html.

[7] Team Hunyuan3D, S. Yang, M. Yang, Y. Feng, X. Huang, S. Zhang, Z. He, D. Luo, H. Liu, Y. Zhao, Q. Lin, Z. Lai, X. Yang, H. Shi, Z. Zhao, B. Zhang, H. Yan, L. Wang, S. Liu, J. Zhang, M. Chen, L. Dong, Y. Jia, Y. Cai, J. Yu, Y. Tang, D. Guo, J. Yu, H. Zhang, Z. Ye, P. He, R. Wu, S. Wei, C. Zhang, Y. Tan, Y. Sun, L. Niu, S. Huang, B. Zheng, S. Liu, S. Chen, X. Yuan, X. Yang, K. Liu, J. Zhu, P. Chen, T. Liu, D. Wang, Y. Liu, Linus, J. Jiang, J. Huang, and C. Guo. Hunyuan3D 2.1: From images to high-fidelity 3D assets with production-ready PBR material. *arXiv preprint arXiv:2506.15442*, 2025. URL https://arxiv.org/abs/2506.15442.

[8] B. Wen, W. Yang, J. Kautz, and S. Birchfield. FoundationPose: Unified 6D pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17868–17879, June 2024. doi:10.1109/CVPR52733.2024.01692.

[9] OpenAI. GPT-6 Astra system card. https://deploymentsafety.openai.com/gpt-6-astra, Sept. 2026. Accessed 2026-10-27.

[10] Anthropic. Introducing the Model Context Protocol. https://www.anthropic.com/news/model-context-protocol, Nov. 2024. Accessed 2026-10-27.

[11] Model Context Protocol. Model Context Protocol specification, version 2026-07-28. https://modelcontextprotocol.io/specification/2026-07-28, 2026. Accessed 2026-10-27.

[12] K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, M. Y. Galliker, D. Ghosh, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, D. LeBlanc, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, A. Z. Ren, L. X. Shi, L. Smith, J. T. Springenberg, K. Stachowicz, J. Tanner, Q. Vuong, H. Walke, A. Walling, H. Wang, L. Yu, and U. Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In J. Lim, S. Song, and H.-W. Park, editors, *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pages 17–40. PMLR, 2025. URL https://proceedings.mlr.press/v305/black25a.html.

[13] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. Abbeel. GELLO: A general, low-cost, and intuitive teleoperation framework for robot manipulators. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12156–12163, 2024. doi:10.1109/IROS58592.2024.10801581.

[14] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024. doi:10.15607/RSS.2024.XX.045.