WORLDMIRROR: AGENTIC REAL-TO-SIM GENERATION OF INTERACTIVE WORLDS FROM IN-THE-WILD ROBOT DEMONSTRATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing approaches for robot learning in simulation often require substantial manual effort to construct scenes and specify tasks, limiting training scalability and data diversity. Meanwhile, in-the-wild robot demonstrations naturally provide diverse and realistic observations of real-world interactions. Leveraging these demonstrations for simulation requires both high-fidelity visual reconstruction and physically consistent interactions. General-purpose agents offer flexible adaptation to diverse scenarios and strong reasoning about 3D spatial relationships, while specialized generative models provide high-fidelity reconstruction of scene appearance and geometry. We introduce **WorldMirror**, an agentic real-to-sim framework that constructs interactive digital twins from in-the-wild robot demonstrations. Our central idea is to formulate real-to-sim construction as the iterative refinement of an executable world program. A coding agent decomposes the observed scene, plans reconstruction steps, and orchestrates specialized image and 3D generation tools, refining the program through verification feedback. A camera-pose-anchored scene representation organizes the world into static visual context, support geometry, and editable interactive assets while preserving spatial relationships among the scene, objects, and robot. The agent uses this representation to diagnose and correct reconstruction errors, enabling physically consistent interaction and trajectory replay. Experiments on DROID demonstrate that WorldMirror can reconstruct visually faithful and geometrically consistent interactive worlds while preserving the physical validity of recorded robot trajectories.

1 INTRODUCTION

Scaling generalist robot policies requires diverse training data across a wide range of real-world scenarios. (Physical Intelligence, 2025; Black et al., 2024; Wang et al., 2026). Yet collecting real-world demonstrations and repeatedly testing policies across environments remain costly and labor-intensive. Simulation offers a complementary route to scalable data generation and controlled, repeatable evaluation (Li et al., 2024; Jain et al., 2025). Realizing this potential requires interactive environments that reproduce the appearance and physical interactions of real-world deployment settings.

For simulation to serve these purposes, scene geometry, object placement, camera viewpoints, and robot configuration must closely match the real setup. Even small errors in scale or contact geometry can alter task outcomes. Recent real-to-sim systems have advanced scene reconstruction and its use in robot learning and evaluation (Zhao et al., 2025; Ranawaka et al., 2026; Jain et al., 2025; Zhang et al., 2026; Dai et al., 2024; Torne et al., 2024). However, reliable construction requires decisions beyond feedforward prediction and parameter optimization. A missing object or unsuitable collision geometry may require revisiting scene decomposition, selecting a different reconstruction tool, or modifying assembly code. These challenges motivate *agentic world modeling*, in which a coding agent plans reconstruction, coordinates tools, and iteratively refines an executable world using verification feedback.

We introduce **WorldMirror**, an agentic real-to-sim framework that converts robot videos and their associated trajectories into interactive digital twins. At its core, a coding agent maintains an ex-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: WorldMirror enables simulation-ready digital twins with strong realism and interaction fidelity. While GPT-6 Astra benefits from large-scale priors for geometry and layout understanding, WorldMirror achieves better visual realism and physical executability through agentic reconstruction and simulator grounding.

ecutable world program, decomposes the scene, selects reconstruction tools, and revises generated assets and assembly code in response to verification feedback. A camera-pose-anchored scene representation organizes the reconstructed world into static visual context, support geometry, and editable interactive assets, while placing all components in a shared coordinate system. This representation preserves the spatial relationships required to replay recorded robot trajectories without modification. Table 1 situates WorldMirror among representative real-to-sim and agentic world construction systems and highlights the capabilities enabled by its design.

We evaluate WorldMirror along three dimensions: visual quality, layout fidelity, and motion replay. On DROID, we assess visual quality, object layout fidelity, and the accuracy of replaying recorded robot trajectories in the reconstructed scenes (DROID Team et al., 2024).

Our contributions are:

- We propose **WorldMirror**, an agentic real-to-sim framework that constructs interactive digital twins through tool orchestration and verification-guided refinement of executable world programs.
- We introduce a camera-pose-anchored scene representation that organizes reconstructed worlds into static visual context, support geometry, and editable interactive assets, enabling spatially consistent replay of recorded robot trajectories.
- We systematically evaluate WorldMirror on DROID across visual quality, layout fidelity, and motion replay, demonstrating its ability to reconstruct visually faithful and spatially consistent interactive worlds.

2 Related Work

Real-to-sim world construction. Constructing interactive digital twins requires recovering both the appearance of a real scene and the physical structure needed to support robot interaction. Existing systems approach this problem through different reconstruction strategies: RialTo builds scanned digital twins for policy improvement (Torne et al., 2024), while RoLA constructs interactive environments from individual images (Zhao et al., 2025). PolaRiS further develops scalable real-to-sim

Table 1: Comparison of representative real-to-sim and agentic world construction systems. Capabilities are categorized into scene reconstruction, robot-world alignment, and agentic construction, covering both world modeling fidelity and autonomous construction mechanisms.

Capability	**RoLA**	**RoboSnap**	**PolaRiS**	**SimFoundry**	**SAGE**	**Astra**	**WorldMirror**
Scene Reconstruction							
Background	✓	✓	✓	✓	✓	✗	✓
Objects	✓	✓	✓	✓	✓	✓	✓
Interactive	✓	✓	✓	✓	✓	✓	✓
Robot Alignment							
Camera	✓	✓	✓	✓	✗	✓	✓
Trajectory Replay	✗	✓	✗	✗	✗	✓	✓
Agentic Construction							
Tool Orchestration	✗	✗	✗	✗	✓	✓	✓
Verification	✗	✗	✗	✓	✓	✓	✓

reconstruction for generalist robot policy evaluation, using neural reconstruction to turn short video scans of real-world scenes into interactive simulation environments (Jain et al., 2025). Several methods further integrate reconstruction with explicit alignment and physical refinement. SimFoundry combines video-based reconstruction, pose alignment, and physics annotation, and generates digital cousins for robot evaluation and training (Ranawaka et al., 2026). RoboSnap separates interactive geometry from Gaussian visual context and refines scene configurations through geometric and physics optimization (Zhang et al., 2026), while TwinAligner jointly addresses visual alignment and rigid-body system identification (Fan et al., 2025). These works establish layered representations and physical alignment as important ingredients of interactive world construction. WorldMirror builds on this foundation, emphasizing their coordination through editable world programs and agent-directed verification.

Agentic 3D modeling. Recent advances in multimodal agents have extended agent capabilities toward programmatic 3D and world construction. GPT-6 Astra (OpenAI, 2026) demonstrates the use of general-purpose agents for constructing and manipulating complex 3D worlds from visual inputs. More specialized systems develop agentic workflows for programmatic 3D modeling. Programmatic 3D modeling enables agents to construct, execute, inspect, and revise 3D assets through code. Articraft develops articulated assets using a domain-specific SDK and execution-based validation (Zhou et al., 2026), while Procedura combines procedural assembly with geometric constraints and visual critique (Lin et al., 2026). SAGE introduces an agentic pipeline that coordinates multiple scene and object generators with critics to iteratively construct simulator-ready environments (Xia et al., 2026). LiteReality reconstructs RGB-D scans into graphics-ready scenes with structured layouts, retrieved object assets, realistic materials, and basic physical properties (**?**). Agentic Real2Sim extends agentic reconstruction to recorded robot interactions, combining object reconstruction, scene calibration, and simulator-guided refinement, with evaluation centered on episode replay (Chen et al., 2026). WorldMirror focuses on reconstructing the deployment environment needed for policy evaluation and learning. Its scope includes complex background appearance and support geometry, interactive object geometry and placement, robot configuration, and the extrinsics and fields of view of external and wrist-mounted cameras. A coding agent coordinates these components within an executable world program and uses verification feedback to revise both individual assets and their assembly.

Simulation for Robot Learning Simulation environments provide an important foundation for scalable robot learning. RLBench offers a diverse suite of language-conditioned manipulation tasks with reusable simulation environments and demonstrations (James et al., 2020), while robosuite provides a modular framework for constructing and evaluating robot manipulation tasks (Zhu et al., 2020). ManiSkill further scales simulation-based learning with diverse manipulation tasks, object assets, and large-scale demonstrations (Gu et al., 2023). SIMPLER demonstrates the use of sim-

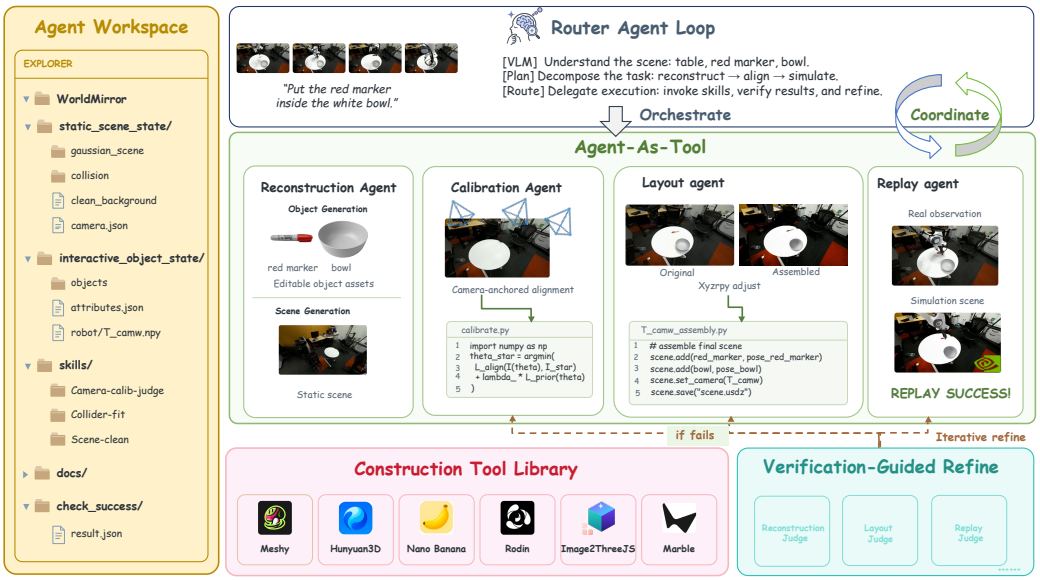

Figure 2: **WorldMirror framework.** A coding agent maintains an editable world, invokes scene and object generation tools, and aligns camera, assets, and robot through assembly code. Rendered observations and execution feedback guide targeted revisions. The resulting world supports trajectory replay, policy evaluation, and simulation-data generation.

ulation as a scalable and reproducible complement to real-world evaluation, providing simulated environments for common real robot setups (Li et al., 2024). MimicGen adapts a small set of human demonstrations to diverse simulated scenes and object configurations (Mandlekar et al., 2023), while RoboTwin provides large-scale simulation environments and task variations for robot manipulation learning (Chen et al., 2025). These approaches demonstrate the potential of simulation for scalable robot learning by providing increasingly diverse simulation environments and large-scale demonstrations. WorldMirror complements these efforts by combining agentic real-to-sim reconstruction with in-the-wild robot demonstrations to construct executable interactive worlds.

3 Method

3.1 Problem Formulation

We study the problem of constructing simulation-ready digital twins from in-the-wild robot manipulation demonstrations. Unlike conventional reconstruction pipelines that formulate real-to-sim conversion as a fixed sequence of perception and generation modules, we view digital twin construction as an agentic world-building process that requires adaptive reasoning, tool selection, and iterative refinement.

Given an RGB video sequence $\mathcal{V}$ and the corresponding robot trajectory $\mathcal{T}$, an agent constructs a digital twin through a sequence of decisions:

$$\hat{\mathcal{S}} = \mathcal{A}(\mathcal{V}, \mathcal{T}),$$

where $\mathcal{A}$ denotes an agentic reconstruction process that dynamically selects reconstruction skills, updates the world state, and verifies intermediate results.

A simulation-ready digital twin should satisfy three requirements. First, it should preserve the visual appearance and spatial organization of the observed environment. Second, task-relevant objects should be represented as editable simulation entities with appropriate geometry, physical properties, and interaction semantics. Third, all reconstructed components, including the scene, objects, and robot embodiment, should share a consistent coordinate system such that the demonstrated trajectory can be replayed without modification.

To enable such construction, WorldMirror maintains an explicit world state containing reconstructed entities, spatial relationships, and intermediate verification results. The final digital twin $\hat{S}$ consists of a reconstructed environment, editable interactive objects, a calibrated camera configuration, and the robot setup required for trajectory replay.

3.2 Structured World State Memory

Agentic world construction requires a structured representation that maintains the evolving status of the reconstructed environment and supports long-horizon reasoning. Unlike conventional real-to-sim pipelines that directly transform observations into final simulation assets, WorldMirror maintains an explicit world state that is iteratively updated throughout the construction process. This state serves as the external memory of the agent, storing reconstructed entities, their spatial relationships, and simulation-relevant attributes required for subsequent skill execution and verification.

Formally, we represent the current reconstruction state as

$$\mathcal{M}_t = (\mathcal{S}_t, \mathcal{O}_t, \mathcal{R}_t),$$

where $\mathcal{S}_t$ denotes the scene state, $\mathcal{O}_t$ denotes the object state, and $\mathcal{R}_t$ captures relational information among reconstructed entities. The state evolves as the agent invokes different reconstruction skills and receives verification feedback.

Static scene state. The static scene state maintains the persistent environmental context of the target task, including fixed elements such as tables, walls, and other background structures. We represent their visual appearance using a Gaussian scene representation and model their physical support through corresponding collision geometry. This state provides the visual and geometric foundation for placing interactive objects and reproducing robot interactions.

Interactive object state. The interactive object state maintains task-relevant entities that participate in robot manipulation. Each object is represented by its geometry, pose, spatial layout, physical properties, and task-relevant interaction attributes, enabling its instantiation as an editable simulation entity. These object states can be selectively updated by reconstruction skills according to the requirements of the target task.

3.3 Agentic World Construction

Router-based Orchestration. WorldMirror performs adaptive world construction through a router agent that reasons over the robot demonstration and current world state. The router decomposes the construction objective into subtasks and dynamically selects appropriate agents and tools, avoiding a fixed reconstruction sequence.

Agent-as-Tool Delegation. WorldMirror adopts an agent-as-tool paradigm, where specialized agents are exposed as callable capabilities to the router agent. These agents operate on the shared world state and provide dedicated capabilities for camera calibration, spatial layout reasoning, asset reconstruction, and trajectory replay, enabling flexible composition of heterogeneous construction operations.

Construction Skill Library. WorldMirror maintains a library of reusable construction skills that provide modular capabilities for digital twin construction. Each skill encapsulates a specific perception, generation, or simulation procedure and can be dynamically invoked by specialized agents according to the current world state. This design allows different agents to compose heterogeneous capabilities rather than relying on a fixed reconstruction pipeline.

- **Camera calibration.** This skill establishes the geometric reference between the real observation and the reconstructed world by estimating camera parameters that align rendered and observed views. We initialize camera correspondence through DINO-based image matching (Oquab et al., 2023) and refine the estimation using structural edge alignment (Von Gioi et al., 2012).

- **Spatial layout.** This skill reasons about the spatial organization of reconstructed entities and determines their placement within the scene. It aligns generated assets with the observed environment and maintains geometric consistency among the static scene, interactive objects, and robot embodiment.
- **Asset construction.** This skill generates simulation-ready visual and physical assets from the reconstructed world state. For static environments, it leverages generative scene reconstruction to recover visual context and supporting geometry, while task-relevant objects are reconstructed as editable assets using appropriate generation methods, such as Meshy for rigid objects and Articraft (Zhou et al., 2026) for articulated objects.
- **Simulation preparation.** This skill converts reconstructed components into simulation-ready entities by configuring physical attributes, interaction properties, and robot execution settings. It provides the necessary setup for downstream trajectory replay and interaction verification.

3.4 Verification-driven Refinement

Consistency Verification. After each construction step, WorldMirror evaluates the consistency of the updated world state with respect to the original demonstration. The verification process examines the rendered observation, reconstructed geometry, and simulated interaction state to identify discrepancies in the current digital twin.

Specifically, visual consistency is evaluated by comparing rendered views with the original observations, while spatial consistency checks the alignment of reconstructed entities under the camera-anchored coordinate system. For interactive tasks, physical consistency is further assessed by examining whether reconstructed object properties and robot trajectories preserve the demonstrated interaction constraints.

Closed-loop Refinement. The verification results are converted into refinement signals and propagated back to the router agent through the world state. Based on the detected inconsistency, the router agent selects the corresponding agents and skills to update the affected components rather than reconstructing the entire scene. For example, camera-related errors trigger calibration refinement, spatial misalignment invokes layout adjustment, while incorrect asset properties lead to object reconstruction updates.

This closed-loop process enables WorldMirror to iteratively improve the digital twin through repeated cycles of construction, verification, and targeted refinement.

3.5 Camera Pose-Anchored Coordinate System

After camera calibration, WorldMirror establishes a camera-anchored coordinate system that serves as the geometric reference for the reconstructed digital twin. Unlike independently reconstructed components that may suffer from inconsistent poses, all scene elements, interactive objects, and robot embodiments are expressed within the same viewer-world frame, preserving the spatial correspondence observed in the original demonstration.

Let

$$T_{\mathrm{camw}} = \begin{bmatrix} R & t \\ \mathbf{0}^\top & 1 \end{bmatrix} \in SE(3)$$

denote the transformation from the camera frame to the unified world frame. All reconstructed components are transformed into this reference frame through rigid transformations:

$$X_w = T_{\mathrm{camw}} X_c,$$

where X_c and X_w denote points represented in the camera and viewer-world coordinate systems, respectively.

This camera-anchored formulation enables consistent assembly of the reconstructed environment, editable object assets, and robot trajectory. Specifically, generated assets are placed according to their observed spatial correspondence, while the demonstrated robot trajectory is transformed into the same coordinate frame without modification. The resulting digital twin preserves both visual alignment and geometric consistency required for downstream simulation and interaction replay.

Table 2: Visual quality comparison on 10 DROID scenes. Higher is better for ↑ metrics, while lower is better for ↓ metrics.

Type	Method	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓	Aesthetic ↑
Open-source	RoLA (Zhao et al., 2025)	13.53	**0.587**	0.506	0.206	0.480
	RoboSnap (Zhang et al., 2026)	10.30	0.483	0.641	0.209	0.532
Proprietary	GPT-6 Astra	11.90	0.515	0.606	0.299	0.472
Ours	**WorldMirror**	**13.57**	0.570	**0.427**	**0.168**	**0.570**

4 EXPERIMENTS

We evaluate WorldMirror on in-the-wild robot manipulation demonstrations from DROID (DROID Team et al., 2024). Given recorded RGB observations and robot trajectories, WorldMirror automatically constructs simulation-ready digital twins through its agentic construction framework. We evaluate the resulting digital twins in terms of visual fidelity, spatial consistency, and interaction consistency against the original demonstrations.

4.1 EXPERIMENTAL SETUP

Comparisons. Our reconstruction baselines are RoLA (Zhao et al., 2025), RoboSnap (Zhang et al., 2026), SimFoundry (Ranawaka et al., 2026), and GPT-6 Astra (OpenAI, 2026). Each table lists the methods evaluated for the corresponding metrics.

Metrics. We evaluate digital twin fidelity at three levels: visual fidelity, spatial fidelity, and replay fidelity. Visual fidelity is measured using PSNR, SSIM, LPIPS, DISTS, and aesthetic scores; spatial fidelity is evaluated using mask IoU, position error, and relative vector and distance errors; and replay fidelity is measured using object-motion errors, contact F1, and success rate.

4.2 MIRRORING REAL-WORLD APPEARANCE

We compare first-frame simulation renders with real exterior-camera images on ten DROID scenes. Table 2 reports quantitative results, and Figure 3 provides qualitative comparisons of reconstructed views. We further visualize the wrist-view reconstruction timeline in Figure 4, showing the correspondence between the observed robot demonstration and the reconstructed simulation sequence. WorldMirror achieves the lowest LPIPS and DISTS, together with competitive PSNR and aesthetic scores among evaluated methods. These results support improved visual correspondence on the evaluated views, particularly under perceptual metrics.

4.3 PRESERVING REAL-WORLD SPATIAL STRUCTURE

To assess whether reconstructed scenes preserve real-world spatial structure, we evaluate object placement and inter-object relationships on a shared set of 17 task-relevant objects and eight object pairs from DROID. We report mask IoU, position error, and relative vector and distance errors, with the 3D metrics computed from visible-surface centroids. As shown in Table 3, WorldMirror outperforms the evaluated open-source baselines across all four metrics. Compared with GPT-6 Astra, however, it has lower mask IoU and higher position and relative vector errors. WorldMirror achieves the lowest relative distance error, indicating better preservation of pairwise distances on the evaluated scenes, but not a consistent advantage across spatial metrics.

4.4 PRESERVING REAL-WORLD MOTION AND INTERACTION

To evaluate whether reconstructed digital twins preserve the physical conditions required for real-world manipulation, we replay the original DROID robot trajectories in the reconstructed environments using Isaac Sim without any trajectory modification or additional motion retargeting. Unlike approaches that require re-planning or inverse kinematics based adaptation to transfer demonstra-

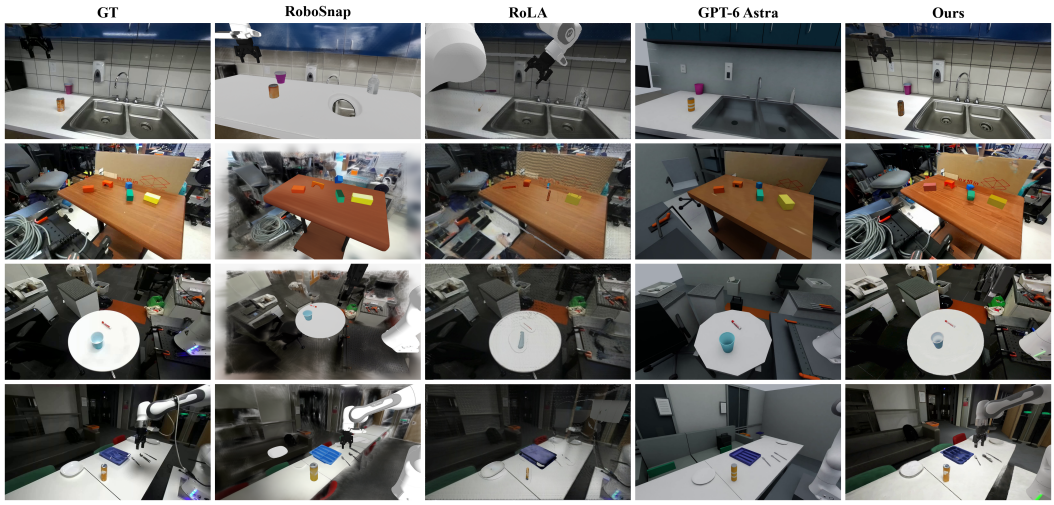

Figure 3: Qualitative RGB comparison with existing baselines on DROID scenes. WorldMirror produces more faithful visual reconstructions with improved appearance consistency and scene details compared with open-source and proprietary baselines.

Table 3: Object layout fidelity comparison. Higher is better for ↑ metrics, while lower is better for ↓ metrics.

Category	Method	IoU ↑	Pos. Err. ↓	Rel. Vec. ↓	Rel. Dist. ↓
Open-source	RoLA (Zhao et al., 2025)	0.232	30.65	0.814	0.542
	RoboSnap (Zhang et al., 2026)	0.313	27.79	0.644	0.583
	SimFoundry (Ranawaka et al., 2026)	0.169	19.54	0.627	0.507
Proprietary	GPT-6 Astra (OpenAI, 2026)	**0.598**	**8.72**	**0.552**	0.331
Ours	**RoboClone**	0.469	9.66	0.557	**0.219**

tions into simulation, WorldMirror directly executes the recorded robot trajectories in the reconstructed coordinate system, providing a stricter evaluation of geometric and physical consistency.

We assess motion consistency using displacement, final-position, direction, and DTW errors, and evaluate interaction fidelity using contact F1 and replay success rate.

As shown in Table 4, WorldMirror achieves the lowest errors across all four motion metrics and higher contact F1 than GPT-6 Astra. It achieves a replay success rate of 100% on the evaluated demonstrations, compared with 40% for GPT-6 Astra. These results demonstrate that the reconstructed digital twins preserve the spatial and physical consistency required for direct trajectory replay and robot-environment interaction.

5 Conclusion

We introduced **WorldMirror**, an agentic real-to-sim framework that constructs simulation-ready digital twins from in-the-wild robot demonstrations. Unlike conventional reconstruction pipelines that rely on fixed sequences of specialized modules, WorldMirror formulates digital twin construction as an adaptive process in which a router agent coordinates specialized agents and skills, maintains a structured world state, and refines the reconstruction through verification-driven feedback.

Experiments on DROID demonstrate that WorldMirror preserves visual appearance, spatial structure, and motion-interaction consistency of real-world demonstrations. By maintaining a camera-pose anchored coordinate system, WorldMirror enables direct trajectory replay in reconstructed en-



Figure 4: Real-to-sim wrist-view comparison over time. The reconstructed digital twin closely matches the real robot observations throughout the manipulation trajectory, demonstrating consistent visual alignment during replay.

Table 4: Motion replay and physical interaction evaluation of reconstructed digital twins on DROID demonstrations. Motion metrics measure trajectory consistency, while interaction metrics evaluate contact fidelity and replay success in Isaac Sim. F1 denotes contact F1 score, and SR denotes replay success rate.

	Motion Replay				Physical Interaction	
Method	Disp. (cm)↓	Final (cm)↓	Dir. (°)↓	DTW (cm)↓	F1↑	SR↑
RoLA (Zhao et al., 2025)	24.70	40.80	53.00	10.30	–	–
GPT-6 Astra (OpenAI, 2026)	11.17	21.98	81.30	10.82	0.653	40%
RoboClone	**1.79**	**5.83**	**10.90**	**1.64**	**0.873**	**100%**

vironments without trajectory modification, providing simulation-ready digital twins for robot interaction evaluation.

We hope WorldMirror serves as a step toward scalable robot learning and evaluation by reducing the manual effort required to transform diverse real-world experiences into reliable simulation environments.

REFERENCES

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Guanxiong Chen, Qianjun Xia, Jiawei Peng, Heng Zhang, Bole Ma, Justin Qian, Ziyi Jiao, Bingyang Zhou, Luoxin Ye, Kaifeng Zhang, et al. Agentic real2sim: Physics-based world modeling with vision-language agents. *arXiv preprint arXiv:2607.19190*, 2026.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu 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.

Tianyuan Dai, Josiah Wong, Yunfan Jiang, Chen Wang, Cem Gokmen, Ruohan Zhang, Jiajun Wu, and Li Fei-Fei. Automated creation of digital cousins for robust policy learning. In *Conference on Robot Learning*, 2024.

DROID Team, Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, et al. DROID: A large-scale in-the-wild robot manipulation dataset. In *Proceedings of Robotics: Science and Systems*, 2024.

Hao Fan, Haoqiang Dai, Jiayuan Zhang, Jiahui Li, Qiang Yan, Yuhang Zhao, Mingyu Gao, Jieming Wu, Hao Tang, and Hao Dong. TwinAligner: Visual-dynamic alignment empowers physics-aware Real2Sim2Real for robotic manipulation. *arXiv preprint arXiv:2512.19390*, 2025.

Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, et al. Maniskill2: A unified benchmark for generalizable manipulation skills. *arXiv preprint arXiv:2302.04659*, 2023.

Arhan Jain, Mingtong Zhang, Kanav Arora, William Chen, Marcel Torne, Muhammad Zubair Irshad, Sergey Zakharov, Yue Wang, Sergey Levine, Chelsea Finn, Wei-Chiu Ma, Dhruv Shah, Abhishek Gupta, and Karl Pertsch. PolaRiS: Scalable real-to-sim evaluations for generalist robot policies. *arXiv preprint arXiv:2512.16881*, 2025.

Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J Davison. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020.

Xuanlin Li, Kyle Hsu, Jiayuan Gu, Oier Mees, Karl Pertsch, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, et al. Evaluating real-world robot manipulation policies in simulation. In *Conference on Robot Learning*, pp. 3705–3728, 2024.

Youtian Lin, Yikang Yang, Zhanpeng Hu, Mengqi Zhou, Feihu Zhang, Xun Cao, Jiaheng Liu, and Yao Yao. Agentic 3D modeling with procedural control. *arXiv preprint*, 2026. URL `https://spatiaos.github.io/projects/procedura/assets/procedura.pdf`.

Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Ikechukwu 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*, 2023.

OpenAI. GPT-6 Astra: A new generation of intelligence. `https://openai.com/index/gpt-6-astra/`, September 2026. Accessed: 2026-09-10.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Physical Intelligence. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Nadun Ranawaka, Josiah Wong, Wei-Lin Pai, Wei-Teng Chu, Tianyuan Dai, Masoud Moghani, Hang Yin, Yunfan Jiang, Wesley Durbano, Brandon Huynh, et al. SimFoundry: Modular and automated scene generation for policy learning and evaluation. *arXiv preprint arXiv:2606.28276*, 2026.

Marcel Torne, Anthony Simeonov, Zechu Li, April Chan, Tao Chen, Abhishek Gupta, and Pulkit Agrawal. Reconciling reality through simulation: A real-to-sim-to-real approach for robust manipulation. In *Proceedings of Robotics: Science and Systems*, 2024.

Rafael Grompone Von Gioi, Jérémie Jakubowicz, Jean-Michel Morel, and Gregory Randall. Lsd: A line segment detector. *Image Processing On Line*, 2:35–55, 2012.

Yuran Wang, Siqiao Huang, Mingleyang Li, Chenhao Zhang, Jiaqi Liang, Weiyang Jin, Yue Chen, Xuemin Chi, Donghao Zhou, Qize Yu, Yu-Kai Wang, Yuhan Rui, Shenzhe Yao, Zhen Yuan, Zhenhao Shen, Kefei Zhu, Zijie Zhu, Ning Gao, Xiaowei Chi, Guanqi He, Shanghang Zhang, Hao Dong, Lin Shao, and Hang Zhao. Openwam: An open, modular exploration towards systematic world-action model pretraining. *arXiv preprint arXiv: 2609.07398*, 2026.

Hongchi Xia, Xuan Li, Zhaoshuo Li, Qianli Ma, Jiashu Xu, Ming-Yu Liu, Yin Cui, Tsung-Yi Lin, Wei-Chiu Ma, Shenlong Wang, et al. Sage: Scalable agentic 3d scene generation for embodied ai. *arXiv preprint arXiv:2602.10116*, 2026.

Shujie Zhang, Jingkun Yi, Weipeng Zhong, Zirui Zhou, Yangkun Zhu, Hanqing Wang, Xudong Xu, Weinan Zhang, and Chunhua Shen. RoboSnap: One-shot real-to-sim scene generation for generalizable robot learning and evaluation. *arXiv preprint arXiv:2607.06699*, 2026.

Sizhe Zhao, Jiafei Mao, Wesley Chow, Zixuan Shangguan, Tiancheng Shi, Ruijie Xue, Yifan Zheng, Yufei Weng, Yang You, Daniel Seita, et al. Robot learning from any images. In *Conference on Robot Learning*, pp. 4226–4245, 2025.

Matt Zhou, Ruining Li, Xiaoyang Lyu, Zhaomou Song, Zhening Huang, Chuanxia Zheng, Christian Rupprecht, Andrea Vedaldi, and Shangzhe Wu. Articraft: An agentic system for scalable articulated 3D asset generation. *arXiv preprint arXiv:2605.15187*, 2026. URL `https://arxiv.org/abs/2605.15187`.

Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Kevin Lin, Abhiram Maddukuri, Soroush Nasiriany, and Yifeng Zhu. robosuite: A modular simulation framework and benchmark for robot learning. *arXiv preprint arXiv:2009.12293*, 2020.