A Real2Sim Digital Twin Pipeline for Photorealistic Robot Simulation: Evaluating VLA Policy Deployment on a Bimanual Mobile Robot

Jasmin Lin, David Keetae Park, Carlos Xavier Soto, Shinjae Yoo, and Wei Xu
Brookhaven National Laboratory, Upton, NY, USA
jaslin5168@gmail.com, {dpark1, csoto, sjyoo, xuw}@bnl.gov

***Abstract*—Digital twins that are automatically constructed from robot sensor data offer a promising pathway for scalable Real2Sim and Sim2Real transfer. However, it remains an open question whether photorealistic reconstruction alone is sufficient to support reliable deployment of vision-language-action (VLA) policies.**

We present a generative-AI-assisted Real2Sim pipeline that generates simulation-ready digital twins from real-world RGB observations with minimal manual intervention. The pipeline uses prompted segmentation to isolate scene components and a generative 3D model to directly produce simulation assets, eliminating the need for traditional multi-view reconstruction or manual 3D modeling.

To evaluate simulation fidelity, we deploy and compare policies from two VLA models in both the real robot and the reconstructed simulation under identical tasks and initial conditions. We compare joint-level action trajectories and analyze how divergence evolves over time in closed-loop execution.

Although the reconstructed environments are visually accurate, we observe increasing trajectory divergence during closed-loop operation. These results indicate that photorealistic reconstruction alone is insufficient to preserve closed-loop control behavior in VLA policies, particularly in contact-rich manipulation settings where small perceptual errors compound over time.

I. Introduction

The ability to automatically construct digital twins from robot sensor data and use those twins for policy learning, evaluation, and transfer is central to scalable embodied AI. Yet, a fundamental concern persists: reconstruction-based Real2Sim pipelines increasingly deliver photorealistic environments, while the embodied policies that must operate within them remain sensitive to even subtle discrepancies in geometry, appearance, dynamics, and sensor characteristics. Understanding when, and to what degree, visual fidelity translates into policy fidelity is therefore a key open problem. Prior work in system identification suggests that accurate dynamics calibration, rather than visual realism alone, is often the dominant factor in successful Sim2Real transfer.

Traditional mitigation strategies such as domain randomization [1], large-scale synthetic data generation, and domain adaptation improve transfer in restricted settings but do not systematically eliminate the structural and dynamic gaps between real and simulated environments. Recent advances in neural scene representations, particularly 3D Gaussian Splatting (3DGS) [2], offer photorealistic rendering with real-time GPU performance, prompting a wave of reconstruction-driven simulation frameworks for robotics [3]–[5]. Concurrently, vision-language-action (VLA) models have emerged as high-capacity, general-purpose policies that integrate perception, language, and action within a single architecture [6]–[8]. However, their sensitivity to distribution shift makes Real2Sim alignment especially critical.

In this work, we evaluate simulation fidelity by comparing VLA policy behavior in real and reconstructed environments under identical tasks and initial conditions. We treat cross-domain consistency of action trajectories as a proxy for simulation fidelity, where divergence reflects mismatches in the reconstructed dynamics.

Our pipeline uses prompted segmentation to separate scene components and a generative 3D model to produce simulation-ready assets for Isaac Sim. This reduces reliance on traditional multi-view reconstruction while still requiring limited user input during segmentation and alignment. These assets are then converted and imported into NVIDIA Isaac Sim.

Our results show that despite high visual fidelity, trajectory divergence accumulates under closed-loop operation in contact-rich manipulation tasks. These findings motivate evaluating visual reconstruction as a controlled factor within a broader Sim2Real system that includes dynamics and embodiment considerations.

This paper makes three contributions: (1) a generative-AI-assisted pipeline for generating simulation-ready digital twins from RGB observations, (2) a cross-domain evaluation protocol based on closed-loop trajectory consistency for policies of two finetuned VLA models, and (3) an empirical analysis showing that photorealistic reconstruction alone is inadequate for consistent policy transfer in contact-rich manipulation tasks.

II. Related Work

A. Real2Sim Pipelines from Multimodal Robot Data

Classical robotics simulation environments rely on CAD tools and manual asset preparation, creating bottlenecks that limit iteration speed and accessibility. Photogrammetry-based pipelines using COLMAP [9], [10] reconstruct geometry from multi-view imagery but still require significant manual cleanup before simulation use. Scalable real-to-sim approaches have automated asset creation by extracting object geometry and

physical parameters from robot interaction data, enabling pick-and-place pipelines without manual intervention [11]. Our work reduces the need for manual 3D modeling, though it still requires limited user input for segmentation and alignment.

B. 3D Gaussian Splatting for Robotic Simulation

3DGS [2] has emerged as a compelling scene representation for robotics due to its photorealistic rendering and real-time GPU performance. RoboGSim [3] integrates 3DGS reconstructions with physics simulators for closed-loop manipulation evaluation. SplatSim [4] replaces mesh-based assets with Gaussian splats to improve zero-shot Sim2Real transfer of RGB manipulation policies. Hybrid representations that couple Gaussian appearance with mesh-based or particle-based geometry for physics interaction are also emerging [12], addressing the critical gap between photorealistic rendering and physically accurate simulation. Our pipeline contributes to this hybrid direction by pairing 3DGS with TRELLIS-decoded meshes, explicitly targeting both visual and physical simulation compatibility.

C. Neural 3D Reconstruction and Generative Asset Creation

Learning-based reconstruction approaches such as DUSt3R [13], MASt3R [14], and MUSt3R [15] reduce dependence on classical calibration pipelines by regressing 3D structure directly from images. TRELLIS [16] is a generative image-to-3D model that decodes structured latent representations into multiple output formats (radiance fields, meshes, Gaussian primitives), enabling asset-centric generation from imagery alone. We adopt TRELLIS as the decoding backbone of our pipeline, as its multi-format output directly supports both visual rendering and physics-compatible mesh assets in a single pass.

D. Digital Twins for Sim2Real Transfer

Recent digital twin frameworks increasingly target closed-loop policy transfer. TwinAligner [17] aligns visual and dynamic properties for physics-aware Real2Sim2Real manipulation transfer. VRRobo [18] constructs photorealistic 3DGS-based digital twins for navigation and locomotion, enabling RGB policy transfer. RealMirror [19] integrates 3DGS with teleoperation data for zero-shot Sim2Real manipulation. These systems demonstrate improved transfer performance, but primarily focus on appearance alignment rather than directly examining cross-domain inference behavior under closed-loop operation, presenting a gap our evaluation framework addresses.

E. System Identification and Domain Adaptation for Sim2Real

A large body of work addresses the sim-to-real gap by explicitly modeling uncertainty in simulator parameters or adapting observations across domains. Bayesian system identification approaches such as BayesSim [20] infer a posterior distribution over simulator parameters using likelihood-free inference, enabling posterior-guided domain randomization that improves robustness to dynamics mismatch. More recent approaches leverage differentiable simulators and gradient-based inference, such as Stein Variational Gradient Descent (SVGD)-based methods [21], which infer multimodal parameter distributions using particle-based optimization and multiple-shooting likelihoods to better match real-world trajectories.

Complementary to dynamics calibration, visual domain adaptation methods aim to reduce perceptual discrepancies between real and simulated observations. VR-Goggles [22] translates real-world inputs into the simulation domain at deployment time, improving policy robustness without modifying simulation assets. Large-scale synthetic data generation frameworks such as ProcTHOR [23] and UnrealROX+ [24] instead address the gap through diversity and scale, enabling policies to generalize across a wide range of environments.

Finally, recent work emphasizes the importance of safety and robustness during deployment. [25] proposes integrating safety constraints directly into both training and deployment pipelines, demonstrating that safety-aware training significantly improves real-world performance despite similar simulation metrics.

In contrast to these approaches, our work focuses on evaluating how photorealistic, generative digital twins affect closed-loop policy behavior, rather than explicitly calibrating simulator dynamics or adapting policy inputs. Our results highlight that visual fidelity alone is insufficient, reinforcing the need to integrate these complementary strategies into future digital twin systems.

F. Vision-Language-Action Models

VLA models integrate perception, language reasoning, and action generation into unified architectures trained on large-scale multimodal datasets [6]. GR00T-N1.5 [7] represents the state of the art in this class, combining a vision-language module with a diffusion transformer for continuous action generation, and is designed for generalist embodied control across robot morphologies. $\pi_{0.5}$ [8] is co-trained on a highly diverse mix of data sources (multiple robots, web data, and high-level semantic tasks) and using a built-in hierarchical structure that predicts subtasks before actions. This combination enables strong open-world generalization, allowing it to perform long, multi-step tasks in entirely new environments. A well-documented limitation of VLA models is their sensitivity to visual distribution shift between training and deployment [6], making them a particularly demanding test case for digital twin fidelity.

G. Evaluation Protocols for Real2Sim and Sim2Real

Robust evaluation of embodied policies across real and simulated domains remains an open challenge. Existing benchmarks (LIBERO [26], CALVIN [27], RLBench [28]) are primarily simulation-internal, while real-world evaluation is limited by cost and safety constraints [29], [30]. Recent work emphasizes the need for cross-domain trajectory consistency metrics and closed-loop testing to capture error accumulation [17], [18]. Our work contributes a practical evaluation protocol

that measures joint-level RMSE and temporal divergence, specifically designed for cross-domain assessment of VLA policies within generative digital twins.

H. Simulation Parameter Inference and Domain Calibration

A complementary line of work addresses the sim-to-real gap through explicit identification of simulator parameters. BayesSim [20] formulates domain randomization as Bayesian inference, learning a posterior distribution over simulator parameters that best explains real-world observations. This enables policies to be trained on parameter distributions that are grounded in reality rather than manually specified ranges. More recent approaches leverage differentiable simulation and gradient-based inference, such as Stein Variational Gradient Descent (SVGD) [21], to directly optimize parameter posteriors from real trajectories while capturing multi-modality and uncertainty. These methods demonstrate that accurate dynamics calibration is critical for robust policy transfer, particularly in contact-rich manipulation tasks. Our work differs in that we do not attempt to calibrate simulator parameters, but instead evaluate how far visual reconstruction alone can support cross-domain policy consistency.

These approaches highlight that accurate parameter inference is often more critical than visual realism for successful Sim2Real transfer, particularly in contact-rich settings. Our work complements this line of research by isolating the contribution of visual reconstruction: rather than calibrating simulator parameters, we evaluate how far photorealistic digital twins alone can support cross-domain policy consistency. The observed performance gap suggests that integrating probabilistic system identification methods such as BayesSim and SVGD into generative digital twin pipelines is a necessary direction for achieving robust transfer.

III. Method

A. Overview

Our pipeline constructs a simulation-ready digital twin from real RGB robot observations through three stages: (1) data preparation and generative segmentation, (2) hybrid 3D asset generation via 3DGS and TRELLIS, and (3) simulator integration and cross-domain VLA evaluation. Figure 1 illustrates the full workflow.

The pipeline streamlines twin generation by shifting the manual burden of asset creation toward automated, generative 3D extraction. Rather than relying on exhaustive manual modeling, the framework provides modular design parameters, such as segmentation prompts, asset alignment strategies, and calibration of physical properties that allow researchers to customize and configure complex interaction environments. This modularity turns individual human-in-the-loop interventions into explicit experimental controls. As such, the system establishes a flexible testbed for evaluating the sensitivity of downstream Sim2Real applications, creating opportunities to study the precise impact of simulated variables on closed-loop policy performance.

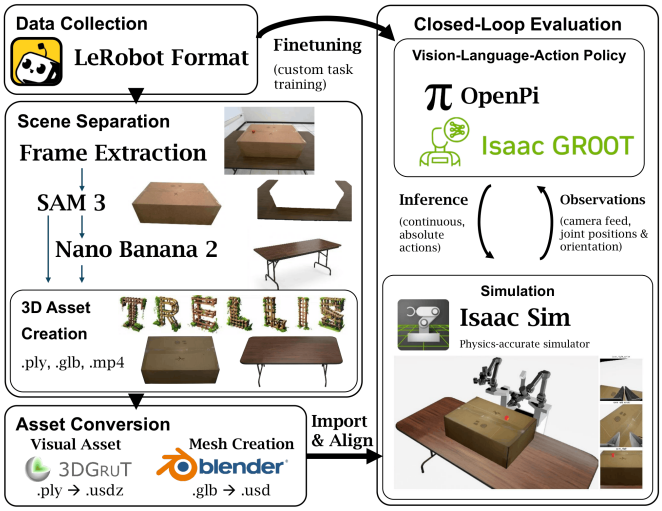

Fig. 1: AI-integrated Real2Sim pipeline. RGB observations are collected from the robot, and scene components are separated using prompted segmentation and regeneration. Isolated object images are passed to a generative 3D model to produce simulation assets. These assets are converted into simulation-compatible formats and imported into Isaac Sim, where they are manually aligned to match the real environment.

B. Robotic Platform

We evaluate our system on a Trossen Mobile AI bimanual mobile manipulation platform equipped with onboard RGB cameras and high-torque actuation. We deploy the Isaac GR00T-N1.5 VLA model [7], a foundation model for generalist embodied control that uses a vision-language perception module and a diffusion transformer for continuous action generation. Deploying GR00T-N1.5 on this platform introduces challenges related to embodiment mismatch, high-torque control dynamics, and sensor characteristics not represented in the model's training distribution. It presents a demanding and realistic test of digital twin fidelity.

C. Data Preparation and Generative Segmentation

We collect high-resolution RGB frames from the robot's onboard cameras during teleoperation, capturing scene geometry, object layouts, and interaction contexts. The first high-resolution frame is passed to a generative segmentation model (e.g., Nano Banana [31], SAM 3 [32]) to decompose the scene into separate object and background layers, enabling modular downstream manipulation of individual scene components. This feature supports scalable scenario generation and environment-level customization within the twin.

D. 3D Asset Generation

Segmented images are passed to TRELLIS [16], a generative image-to-3D model that produces simulation-ready assets. TRELLIS outputs both a mesh representation (`.glb`) and a Gaussian-based representation (`.ply`).

The mesh is used for physics interaction in simulation, while the Gaussian representation provides visual detail. This allows the generated assets to support both rendering and interaction

Fig. 2: Qualitative visual comparison between real environment (left) and reconstructed digital twin (right). The twin preserves key scene features including object layout, geometry, and surface texture.

within Isaac Sim. We pass the reconstructed imagery and scene components to TRELLIS [16], which encodes them into a structured 3D latent representation and decodes into both `.glb` mesh and `.ply` point cloud formats. The mesh is converted into `.usd` format via Blender, and the point cloud is converted via 3dgrut [33], [34] into `.usdz` format. This yields Isaac Sim-friendly mesh and visual assets suitable for collision detection and rigid body physics, directly addressing the gap between photorealistic rendering and physical simulation.

The hybrid pairing of 3DGS (appearance) and meshes (physics) is designed to bridge the dynamics and appearance gap identified as a key challenge for Sim2Real transfer.

E. Simulator Integration

Mesh and Gaussian assets are imported into NVIDIA Isaac Sim. Assets and camera viewpoints are manually aligned in simulation by matching the rendered camera view to the real robot camera feed. A side-by-side of the simulation and real environment is shown in Figure 2.

F. Cross-Domain Evaluation Protocol

We deploy both VLAs across three evaluation domains:

- **Real robot**: Physical execution on the Trossen Mobile AI platform.
- **Isaac Sim digital twin**: Closed-loop simulation within the reconstructed environment.
- **VR interactive twin**: Real-time VR visualization for qualitative inspection and debugging.

For each domain, we record joint-level action trajectories and under identical language instructions and initial conditions. Cross-domain consistency is quantified via *Root Mean Square Error*:

$$\text{MSE}_{\text{norm}}(t) = \sum_{i=1}^{6} w_i \frac{(q_i^{\text{sim}}(t) - q_i^{\text{real}}(t))^2}{(\Delta q_i)^2} + w_g \frac{(g^{\text{sim}}(t) - g^{\text{real}}(t))^2}{(\Delta g)^2},$$

$$\text{RMSE}_{\text{norm}}(t) = \sqrt{\text{MSE}_{\text{norm}}(t)},$$

$$\text{MSE}_{\text{overall}}(t) = \frac{1}{2}\left(\text{MSE}_{\text{left}}(t) + \text{MSE}_{\text{right}}(t)\right),$$

where $q_i^{\text{sim}}(t)$ and $q_i^{\text{real}}(t)$ are the simulated and real joint angles/displacements for joint i at time t, and $g^{\text{sim}}(t)$ and $g^{\text{real}}(t)$ are the scalar gripper actions for the corresponding end-effectors. Δq_i and Δg are the normalization scaling factors for the joints and the gripper, respectively. The parameters w_i and w_g are the weights assigned to the joint and gripper errors.

Also, we adopt temporal divergence profiles using forward kinematics [35], which capture how cross-domain error accumulates over the course of a task.

$$T(q) = \left(\prod_{i=0}^{5} T_i(q_i)\right) T_{5,\text{flange}}\, T_{\text{flange,tool}},$$

$$T_i(q_i) = \begin{bmatrix} \cos\theta_i & -\sin\theta_i \cos\alpha_i & \sin\theta_i \sin\alpha_i & a_i \cos\theta_i \\ \sin\theta_i & \cos\theta_i \cos\alpha_i & -\cos\theta_i \sin\alpha_i & a_i \sin\theta_i \\ 0 & \sin\alpha_i & \cos\alpha_i & d_i \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$p = T(q)_{0:3,3}, \quad R = T(q)_{0:3,0:3},$$

where $T_i(q_i)$ represents the standard Denavit-Hartenberg (DH) transformation matrix for joint variable q_i, defined by the link twist α_i, link length a_i, and link offset d_i. $T_{\text{flange,tool}}$ is the constant transformation from the robot's flange to the tool center point, while p and R are the extracted end-effector position vector and rotation matrix derived from the transformation matrix $T(q)$.

This evaluation protocol is designed to be reusable and environment agnostic, contributing a practical benchmark methodology for digital twin assessment in mobile manipulation.

IV. Experiments

A. Visual Fidelity of the Reconstructed Twin

We qualitatively assess the reconstructed twin by comparing real RGB observations to corresponding rendered views from Isaac Sim. The TRELLIS-generated mesh combined with 3DGS preserves scene layout, object geometry, surface texture, and spatial arrangement.

First, we record 30 episodes for a pick-and-place cube task via teleoperation with a pair of leader arms. The robot has 16 joints, 5 left joints, 1 left gripper, 5 right joints, 1 right gripper, and 2 base joints. All joints with the exception of the grippers (meters) use absolute radians. We exclude the base joints for now because of our stationary tasks. These episodes are used to finetune GR00T-N1.5 and $\pi_{0.5}$ separately. Next, a closed-loop evaluation was run on the VLA's inference to determine whether the model could control the system stably and realistically. The evaluator compares inference on a set of ground truth observations with the ground truth actions. Figures 3 and 4 illustrate the closed-loop evaluation of GR00T-N1.5 and $\pi_{0.5}$, respectively. This step is to ensure that our finetuned VLA has quality inference before we deploy it on the robot.

B. Visual Analysis of VLA Inference

We perform a qualitative comparison of VLA inference behavior in the real robot (Fig. 5) and the reconstructed simulation environment (Fig. 6). Event-level frames are extracted from synchronized task executions under identical language

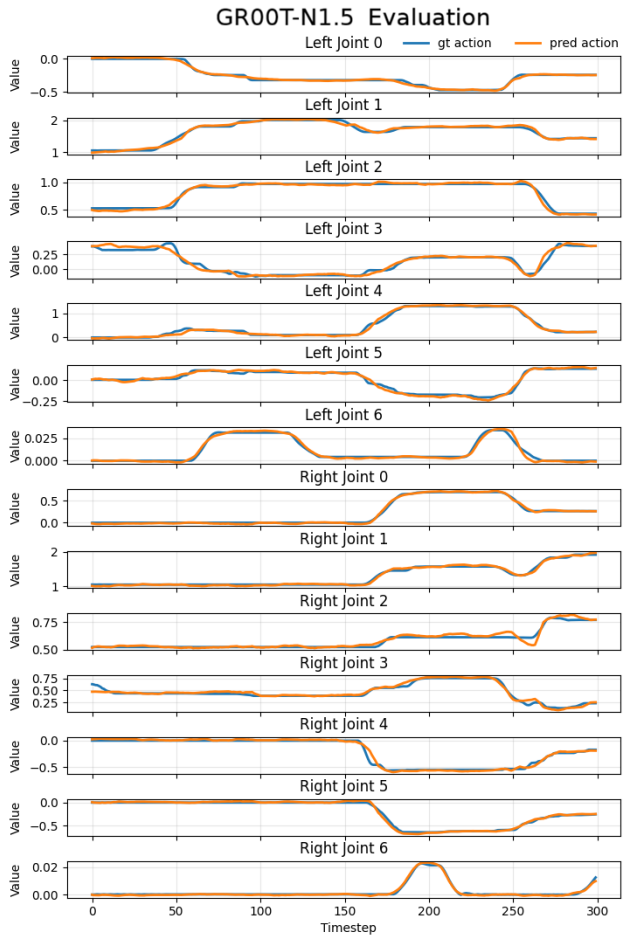

Fig. 3: Closed-loop evaluation graph of GR00T-N1.5 inference plotted alongside ground truth actions.

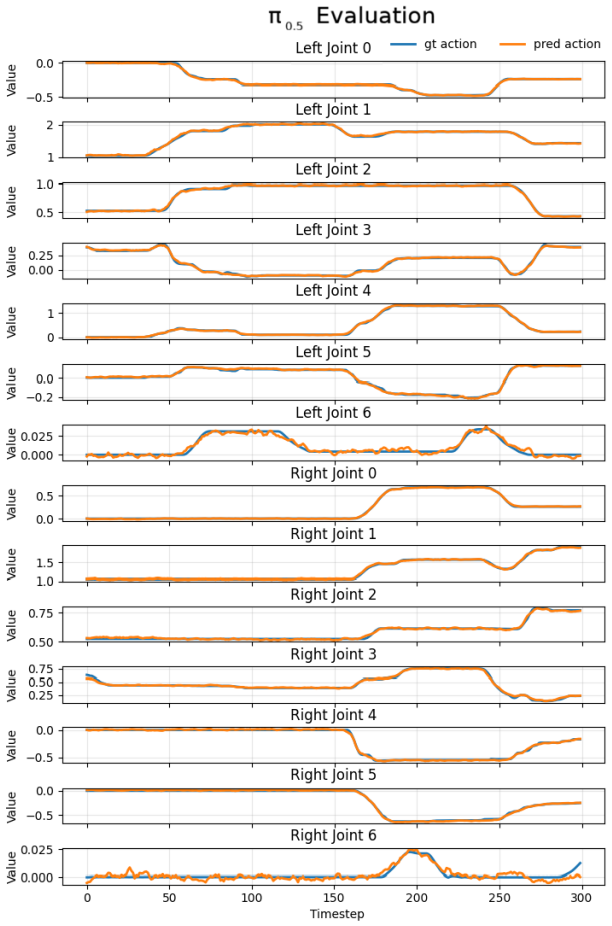

Fig. 4: Closed-loop evaluation graph of $\pi_{0.5}$ inference plotted alongside ground truth actions.

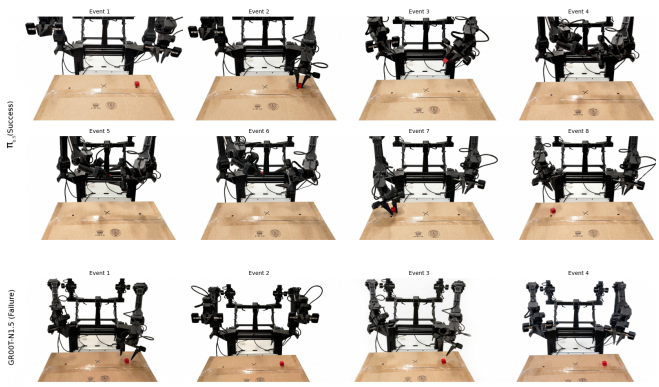

Fig. 5: Event plot depicting key events in real deployment for $\pi_{0.5}$: (1) home position; (2) left grippers grasp; (3) cube pickup; (4) right arm alignment for cube pass; (5) right grippers grasp; (6) left arm release; (7) cube place-down; (8) return to home position. For GR00T-N1.5: (1) home position; (2) left arm finds cube; (3) left arm aligns with cube; (4) left arm fails to grasp cube.

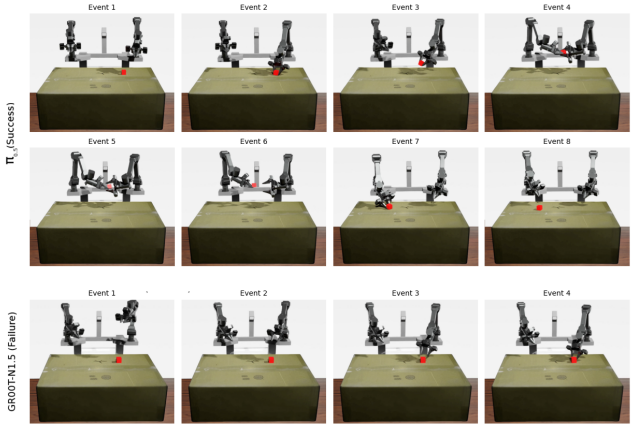

Fig. 6: Event plot depicting key events in simulation deployment for $\pi_{0.5}$: (1) home position; (2) left grippers grasp; (3) cube pickup; (4) right arm alignment for cube pass; (5) right grippers grasp; (6) left arm release; (7) cube place-down; (8) return to home position. For GR00T-N1.5: (1) home position; (2) left arm finds cube; (3) left arm aligns with cube; (4) left arm fails to grasp cube.

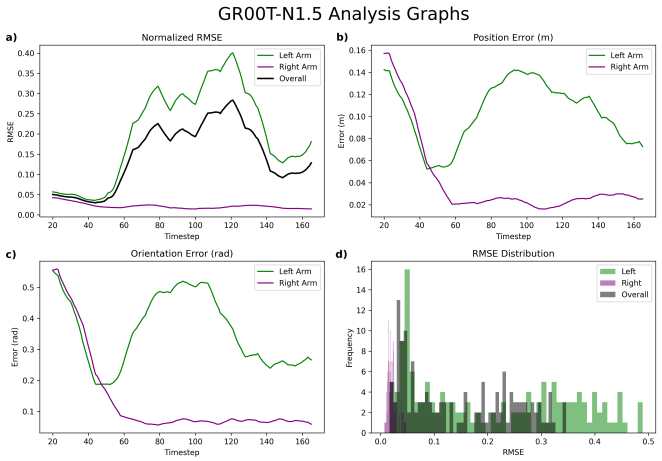

Fig. 7: Analysis graphs for evaluation of GR00T-N1.5, including: (a) RMSE plot of the arms over time including overall RMSE; (b) individual arm position error; (c) individual arm orientation error; and (d) RMSE distribution.

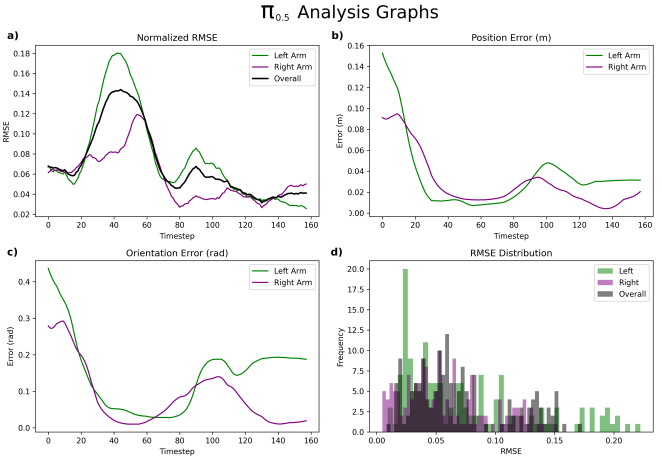

Fig. 8: Analysis graphs for evaluation of $\pi_{0.5}$, including: (a) RMSE plot of the arms over time including overall RMSE; (b) individual arm position error; (c) individual arm orientation error; and (d) RMSE distribution.

instructions and initial conditions. This enables alignment of key decision points in the action sequence and direct comparison of how each policy progresses through the task across domains.

Across both environments, we observe that $\pi_{0.5}$ exhibits consistent phase alignment between real and simulated execution, including grasping, transfer, and placement stages. In contrast, GR00T-N1.5 shows early divergence at the grasping stage, where failure to establish stable contact in the real robot prevents progression to downstream task phases. This failure mode is partially mirrored in simulation, where the policy exhibits repeated hovering behavior over the cube without transitioning into a successful grasp.

These results highlight a key limitation of reconstruction-based digital twins: while event sequences may appear visually aligned at a coarse level, failure modes emerge at contact-rich interaction stages. In particular, discrepancies between real and simulated grasp success indicate that visual alignment alone does not guarantee consistent execution semantics. This supports our broader finding that closed-loop VLA behavior is highly sensitive to embodiment and dynamics mismatch, even under high-fidelity visual reconstruction.

C. Cross-Domain Joint Trajectory Consistency

Figure 7 shows the trajectory error for GR00T-N1.5. The RMSE increases steadily over time, indicating that differences between real and simulated trajectories accumulate during execution. The left arm shows a significant increase in end-effector position and orientation error, while the right arm remains relatively stable. This is because the model fails to complete the grasping stage, preventing the task from progressing to the handover phase.

Figure 8 shows the trajectory error for $\pi_{0.5}$. In this case, RMSE increases during the middle of execution but stabilizes toward the end. The overall error values are lower compared to GR00T-N1.5, and both arms show more balanced behavior. This reflects that $\pi_{0.5}$ is able to execute a more complete sequence of actions, even though differences between real and simulated trajectories are still present.

This temporal divergence pattern is consistent with compounding perception errors under closed-loop operation: minor initial discrepancies in rendered observations produce small action deviations, which alter the robot's state and thus subsequent observations, causing errors to accumulate non-linearly over time. This finding underscores the importance of closed-loop evaluation, as single-step or open-loop metrics would not capture this dynamics gap.

V. Discussion

While qualitative and trajectory-based evaluations provide insight into cross-domain divergence, performance variability across domains remains significant, particularly in contact-rich phases such as grasping and object transfer. This reflects both the inherent difficulty of closed-loop VLA control and the sensitivity of current policies to embodiment and dynamics mismatch. Rather than serving as a full Sim2Real solution, the proposed pipeline isolates the contribution of visual reconstruction within a constrained evaluation setting.

A. Visual Fidelity Is Necessary But Not Sufficient

The negative results observed in our experiments align with a long-standing hypothesis in the Sim2Real literature: that visual fidelity alone is insufficient for reliable policy transfer, particularly in contact-rich manipulation tasks. However, this assumption has rarely been directly tested in the context of modern generative digital twins and high-capacity VLA policies. Our results provide empirical evidence that even highly photorealistic, generatively constructed environments fail to preserve closed-loop policy behavior, thereby isolating visual

reconstruction as a limited method for evaluating Sim2Real readiness.

In contrast, our pipeline focuses on reconstructing environments from visual observations without explicit dynamics calibration. The observed trajectory divergence under closed-loop execution therefore highlights the limitations of relying on photorealistic reconstruction alone, and reinforces the need to integrate parameter inference and dynamics-aware modeling into generative digital twin pipelines.

B. Identified Failure Modes

Systematic analysis across the three evaluation domains identifies the following failure modes:

Perceptual sensitivity. GR00T-N1.5 exhibits high sensitivity to minor rendering discrepancies, i.e., lighting variations, texture smoothing, and boundary artifacts in TRELLIS-decoded meshes, producing disproportionate action deviations from small visual differences.

Embodiment mismatch. GR00T-N1.5 is trained on data from humanoid and other platforms; its action distribution does not fully align with the Trossen Mobile AI's high-torque actuation characteristics, producing control signals that are unstable at contact-rich manipulation phases.

Perception and reconstruction artifacts. The use of generative 3D models such as TRELLIS introduces potential geometric and structural artifacts, including inaccurate object boundaries, surface smoothing, and inconsistencies between visual appearance and physical geometry. These artifacts can degrade both perception (via rendered observations) and interaction (via mesh-based physics), contributing to policy instability. Currently, the pipeline does not explicitly detect or correct such errors, and future work should incorporate validation mechanisms (e.g., multi-view consistency checks or physics-based plausibility constraints) to improve reconstruction reliability.

Closed-loop error compounding. Initial perceptual discrepancies compound over time under closed-loop operation, driving increasing trajectory divergence. This highlights the critical importance of closed-loop, rather than open-loop or single-step, evaluation protocols for digital twin assessment.

Dynamics gap. The current hybrid representation provides physics-compatible meshes for rigid body simulation but does not model contact dynamics, object compliance, or sensor noise at the fidelity needed to match real interaction behavior.

Static scene limitation. The pipeline models the scene at a single point in time and does not support dynamic object tracking or twin updates during task execution, limiting applicability to tasks involving object state changes.

C. Implications for Hybrid Physics-Generative Models

These failure modes collectively point toward the need for hybrid physics-generative digital twin architectures that couple photorealistic generative rendering with accurate dynamics simulation. Gaussian-based appearance could be paired with particle-based or articulated rigid-body physics models to better approximate real contact dynamics. Continuously updated twins refreshed as the robot interacts with the environment would further reduce the gap by keeping the digital twin synchronized with evolving scene state. These directions align directly with the workshop's interest in hybrid physics-generative models and learning algorithms that use continuously updated twins.

D. Safety and Robustness in Closed-Loop Operation

Our closed-loop evaluation also raises safety considerations relevant to digital twin deployment: the compounding error dynamics we observe suggest that policies deployed within imperfect digital twins may exhibit unpredictable behavior that becomes unsafe over extended operation. Uncertainty modeling where the policy or a monitoring layer estimates confidence in the twin's fidelity and degrades gracefully represents a critical direction for safe closed-loop deployment.

VI. Conclusion

In this work, we introduced a generative-AI-assisted Real2Sim pipeline for constructing simulation-ready robotic digital twins and evaluated simulation fidelity by comparing closed-loop VLA policy behavior between the real and the simulation environments through joint-level trajectory analysis. Trajectories are recorded under identical instructions and initial conditions, and divergence is measured over time.

Critically, our study establishes that high visual fidelity, while achievable with the proposed pipeline, does not guarantee robust Sim2Real transfer for modern VLA policies. Identified failure modes, including perceptual sensitivity, embodiment mismatch, closed-loop error compounding, and dynamics gaps, motivate concrete future directions in hybrid physics–generative modeling, continuously updated twins, and robustness-aware learning. The evaluation protocol and failure mode taxonomy we contribute are intended as practical tools for the community working toward reliable generative digital twins in robotic applications.

Acknowledgment

This research was supported by Brookhaven National Laboratory's Lab Directed Research and Development projects 24-063 and 26-052 and the Office of Science of the U.S. Department of Energy (DOE) under FWP CC152 for the Scientific Embodied Agents Lab (SEAL) at Brookhaven National Laboratory.

References

[1] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," 2017. [Online]. Available: https://arxiv.org/abs/1703.06907

[2] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," 2023. [Online]. Available: https://arxiv.org/abs/2308.04079

[3] X. Li, J. Li, Z. Zhang, R. Zhang, F. Jia, T. Wang, H. Fan, K.-K. Tseng, and R. Wang, "Robogsim: A real2sim2real robotic gaussian splatting simulator," 2025. [Online]. Available: https://arxiv.org/abs/2411.11839

[4] M. N. Qureshi, S. Garg, F. Yandun, D. Held, G. Kantor, and A. Silwal, "Splatsim: Zero-shot sim2real transfer of rgb manipulation policies using gaussian splatting," 2024. [Online]. Available: https://arxiv.org/abs/2409.10161

[5] G. Chhablani, S. Bhagat, and D. Mehra, "EmbodiedSplat: Embodied gaussian splatting for robotic manipulation learning," *arXiv preprint*, 2025.

[6] K. Kawaharazuka, J. Oh, J. Yamada, I. Posner, and Y. Zhu, "Vision-language-action models for robotics: A review towards real-world applications," *IEEE Access*, vol. 13, pp. 162 467–162 504, 2025.

[7] NVIDIA, :, J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. J. Fan, Y. Fang, D. Fox, F. Hu, S. Huang, J. Jang, Z. Jiang, J. Kautz, K. Kundalia, L. Lao, Z. Li, Z. Lin, K. Lin, G. Liu, E. Llontop, L. Magne, A. Mandlekar, A. Narayan, S. Nasiriany, S. Reed, Y. L. Tan, G. Wang, Z. Wang, J. Wang, Q. Wang, J. Xiang, Y. Xie, Y. Xu, Z. Xu, S. Ye, Z. Yu, A. Zhang, H. Zhang, Y. Zhao, R. Zheng, and Y. Zhu, "Gr00t n1: An open foundation model for generalist humanoid robots," 2025. [Online]. Available: https://arxiv.org/abs/2503.14734

[8] P. Intelligence, 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," 2025. [Online]. Available: https://arxiv.org/abs/2504.16054

[9] J. L. Schönberger and J.-M. Frahm, "Structure-from-Motion Revisited," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[10] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, "Pixelwise View Selection for Unstructured Multi-View Stereo," in *European Conference on Computer Vision (ECCV)*, 2016.

[11] N. Pfaff, E. Fu, J. Binagia, P. Isola, and R. Tedrake, "Scalable real2sim: Physics-aware asset generation via robotic pick-and-place setups," 2025. [Online]. Available: https://arxiv.org/abs/2503.00370

[12] J. Abou-Chakra, K. Rana, F. Dayoub, and N. Suenderhauf, "Physically embodied gaussian splatting: A visually learnt and physically grounded 3d representation for robotics," in *Proceedings of The 8th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, P. Agrawal, O. Kroemer, and W. Burgard, Eds., vol. 270. PMLR, 06–09 Nov 2025, pp. 513–530. [Online]. Available: https://proceedings.mlr.press/v270/abou-chakra25a.html

[13] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud, "Dust3r: Geometric 3d vision made easy," 2024. [Online]. Available: https://arxiv.org/abs/2312.14132

[14] V. Leroy, Y. Cabon, and J. Revaud, "Grounding image matching in 3d with mast3r," 2024. [Online]. Available: https://arxiv.org/abs/2406.09756

[15] Y. Cabon, L. Stoffl, L. Antsfeld, G. Csurka, B. Chidlovskii, J. Revaud, and V. Leroy, "Must3r: Multi-view network for stereo 3d reconstruction," 2025. [Online]. Available: https://arxiv.org/abs/2503.01661

[16] J. Xiang, Z. Lv, S. Xu, Y. Deng, R. Wang, B. Zhang, D. Chen, X. Tong, and J. Yang, "Structured 3d latents for scalable and versatile 3d generation," 2025. [Online]. Available: https://arxiv.org/abs/2412.01506

[17] H. Fan, H. Dai, J. Zhang, J. Li, Q. Yan, Y. Zhao, M. Gao, J. Wu, H. Tang, and H. Dong, "Twinaligner: Visual-dynamic alignment empowers physics-aware real2sim2real for robotic manipulation," 2025. [Online]. Available: https://arxiv.org/abs/2512.19390

[18] P. Zhu *et al.*, "VRRobo: Photorealistic 3DGS-based digital twins for robot navigation and locomotion," *arXiv preprint*, 2025.

[19] C. Tai, Z. Zheng, H. Long, H. Wu, H. Xiang, Z. Long *et al.*, "RealMirror: A comprehensive, open-source vision-language-action platform for embodied AI," *arXiv preprint arXiv:2509.14687*, 2025.

[20] F. Ramos, R. C. Possas, and D. Fox, "Bayessim: adaptive domain randomization via probabilistic inference for robotics simulators," 2019. [Online]. Available: https://arxiv.org/abs/1906.01728

[21] E. Heiden, C. E. Denniston, D. Millard, F. Ramos, and G. S. Sukhatme, "Probabilistic inference of simulation parameters via parallel differentiable simulation," 2022. [Online]. Available: https://arxiv.org/abs/2109.08815

[22] J. Zhang, L. Tai, P. Yun, Y. Xiong, M. Liu, J. Boedecker, and W. Burgard, "Vr-goggles for robots: Real-to-sim domain adaptation for visual control," 2019. [Online]. Available: https://arxiv.org/abs/1802.00265

[23] M. Deitke, E. VanderBilt, A. Herrasti, L. Weihs, J. Salvador, K. Ehsani, W. Han, E. Kolve, A. Farhadi, A. Kembhavi, and R. Mottaghi, "Procthor: Large-scale embodied ai using procedural generation," 2022. [Online]. Available: https://arxiv.org/abs/2206.06994

[24] P. Martinez-Gonzalez, S. Oprea, J. A. Castro-Vargas, A. Garcia-Garcia, S. Orts-Escolano, J. Garcia-Rodriguez, and M. Vincze, "Unrealrox+: An improved tool for acquiring synthetic data from virtual 3d environments," 2021. [Online]. Available: https://arxiv.org/abs/2104.11776

[25] K. Wrede, S. Zarnack, Y. Di, J. Neumann, M. Dehmel, R. Martin, D. Mayer, and P. Schneider, "Towards a workflow for safe simulation-to-reality transfer of robot control policies," in *2025 13th International Conference on Control, Mechatronics and Automation (ICCMA)*. IEEE, 2025, pp. 175–182.

[26] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone, "Libero: Benchmarking knowledge transfer for lifelong robot learning," 2023. [Online]. Available: https://arxiv.org/abs/2306.03310

[27] O. Mees, L. Hermann, E. Rosete-Beas, and W. Burgard, "Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks," 2022. [Online]. Available: https://arxiv.org/abs/2112.03227

[28] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, "Rlbench: The robot learning benchmark & learning environment," 2019. [Online]. Available: https://arxiv.org/abs/1909.12271

[29] H. Jeong, H. Lee, C. Kim, and S. Shin, "A survey of robot intelligence with large language models," *Applied Sciences*, vol. 14, no. 19, 2024.

[30] J. Wang, Z. Wu, Y. Li, H. Jiang, P. Shu, E. Shi, H. Hu, C. Ma, Y. Liu, X. Wang, Y. Yao, X. Liu, H. Zhao, Z. Liu, H. Dai, L. Zhao, B. Ge, X. Li, T. Liu, and S. Zhang, "Large language models for robotics: Opportunities, challenges, and perspectives," 2024. [Online]. Available: https://arxiv.org/abs/2401.04334

[31] Google DeepMind, "Nano banana," https://deepmind.google/models/gemini-image/, 2025, accessed: 2026-03-22.

[32] 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," 2026. [Online]. Available: https://arxiv.org/abs/2511.16719

[33] N. Moenne-Loccoz, A. Mirzaei, O. Perel, R. de Lutio, J. M. Esturo, G. State, S. Fidler, N. Sharp, and Z. Gojcic, "3d gaussian ray tracing: Fast tracing of particle scenes," *ACM Transactions on Graphics and SIGGRAPH Asia*, 2024.

[34] Q. Wu, J. Martinez Esturo, A. Mirzaei, N. Moenne-Loccoz, and Z. Gojcic, "3dgut: Enabling distorted cameras and secondary rays in gaussian splatting," *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[35] J. Sayono, "6-dof robot arm kinematics (c++)," https://github.com/jacobsayono/6dof-kinematic, 2023, gitHub repository.