

TECHNICAL REPORT

YUBI-Real2Sim: 3DGS Scene Verification for Reliable Policy Evaluation

Hiroki Nishizawa^{1,2} Takehiko Ohkawa¹ Nicola Scianca¹
Tatsuya Matsushima¹ Shigeo Morishima² Kei Ota¹

¹AI Robot Association (AIRoA) ²Waseda University

Abstract

Evaluating VLA policies on real robots is costly, time-consuming, and difficult to reproduce, making real-to-sim simulation essential for scalable evaluation. While 3D Gaussian Splatting (3DGS) enables rapid real-to-sim scene generation, the reconstruction quality required for reliable policy evaluation remains unquantified. In particular, policies relying on wrist-mounted fisheye cameras (e.g., YUBI handheld gripper data) are heavily exposed to visual artifacts (holes, floaters) in wide-angle views, directly degrading task performance. To address this, we present *YUBI-Real2Sim*, a unified framework for 3DGS-based scene reconstruction and real-to-sim policy evaluation under dual fisheye wrist views. We develop an integrated pipeline combining 200-degree fisheye camera capture, 3DGS training, 3D foundation models for object modeling, and MuJoCo-based physical simulation, reproducing the observed real-world policy ordering in eight of nine scene-task comparisons. Using this framework, we construct a benchmark across two bimanual robot platforms, three scenes, three tasks, and dual $\pi_{0.5}$ -based policies. This systematically measures real-to-sim rank agreement and evaluates policy sensitivity against counterfactual 3DGS artifacts (ceiling holes, floaters, warping) and capture disturbances. We further analyze whether novel-view quality scores can indicate policy-ranking reliability under a scene-held-out evaluation protocol. The results show task-dependent performance and limitations of quality-based reliability prediction, highlighting the need to assess reconstruction quality when using Real2Sim for policy evaluation. These results do not yet establish a generally reliable scene-acceptance rule. Project Page: <https://yubi-real2sim.airoa.io>.

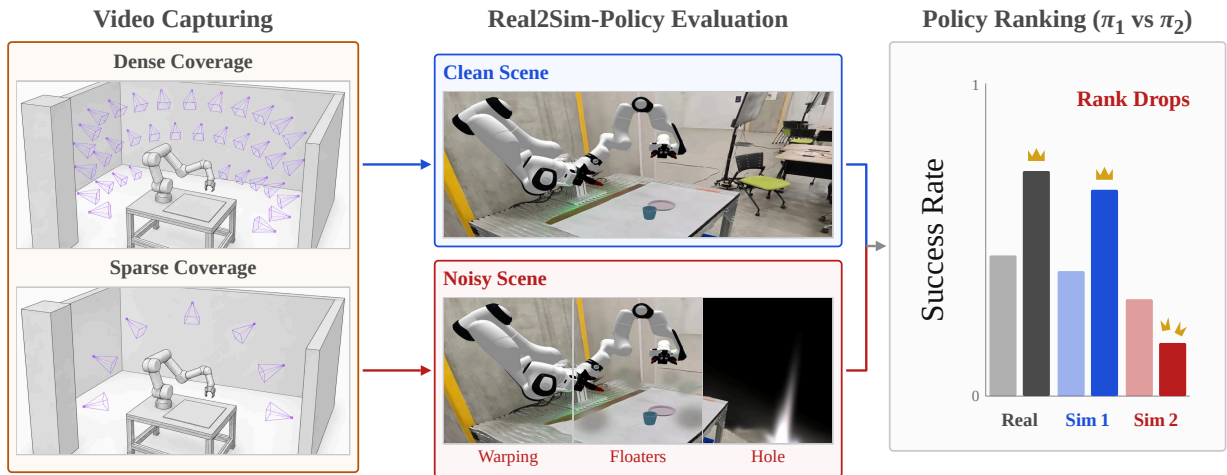


Figure 1 | YUBI-Real2Sim analyzes and verifies 3DGS scenes for reliable policy-ranking evaluation. The quality of 3DGS can easily degrade depending on the camera views. We analyze how variations in 3DGS quality affect policy rankings and examine whether scene-quality scores can indicate policy-ranking reliability.

I. Introduction

Evaluating VLA policies requires repeated robot execution and physical resets, limiting the policies and environments that can be compared [Jin et al. \[2026\]](#). A practical use of simulation is to choose which of two candidate policies to deploy. For that decision, the relevant outcome is whether simulation preserves their real-world policy ranking.

Real-to-sim systems reconstruct environments from real captures and use them for policy evaluation [Jain et al. \[2025\]](#); [Ranawaka et al. \[2026\]](#). 3D Gaussian Splatting (3DGS) [Kerbl et al. \[2023\]](#) makes this approach visually expressive, while SIMPLER [Li et al. \[2025\]](#) motivates judging a simulator by the policy comparisons it supports. However, a reconstructed background may contain holes, floaters, and warped surfaces in regions poorly observed during capture. PolaRiS uses 2D Gaussian Splatting and co-trains policies on real and simulated data to improve policy-ranking correlation [Jain et al. \[2025\]](#).

This issue is particularly relevant to YUBI [Ohkawa et al. \[2026\]](#), a UMI-style [Chi et al. \[2024\]](#) interface for bimanual demonstration collection. YUBI policies observe the environment through two independently moving fisheye wrist cameras. Their wide-angle observations include regions beyond the immediate manipulation target, and their viewpoints change throughout execution. Evaluating these policies therefore requires both a suitable visual simulator and evidence that its scene reconstruction supports reliable policy selection.

We present *YUBI-Real2Sim*, a benchmark and reconstruction pipeline for this setting. It covers Duo and Flat on a dual Franka FR3 setup and the TX-G2 scene on a TX-G2 (AgiBot-G2-compatible) platform, written *TX-G2* throughout. Each task supplies two $\pi_{0.5}$ -based candidate policies [Physical Intelligence \[2025\]](#); one scene-task pair contributes one binary policy comparison.

First we show the clean reconstruction from our pipeline achieves **88.9%** agreement with the real policy rankings in three scenes with three tasks. We then use the clean reconstruction as a reference and edit only the background Gaussians. Ceiling holes, floaters, and warping test whether visual defects cause previously correct policy comparisons to become reversed or tied. Reduced-coverage reconstructions serve a supporting role: they test whether analogous visible defects arise when the input images provide less support. Finally, we evaluate NVS-SQA [Qu et al. \[2026\]](#) as a candidate indicator of ranking reliability, with separate validation required for a scene-rejection rule.

Our contributions are:

- **A Real2Sim pipeline for YUBI policy evaluation.** We integrate scene reconstruction and dual-fisheye policy evaluation on the dual Franka FR3 and TX-G2 platforms. We will release scenes, environments, evaluation code.
- **A benchmark for YUBI policy evaluation in Real2Sim.** We evaluate two $\pi_{0.5}$ -based candidate policies per task across three scenes and three tasks: Cup, Tape, and Pen. Clean simulation reproduces the observed real policy ranking in 8 of 9 scene-task comparisons (88.9%).
- **An analysis of reconstruction defects and policy rankings.** We analyze how counterfactual 3DGS artifacts affect policy rankings under fixed physics, objects, and rollout conditions. We also examine reduced-coverage reconstructions to study defects arising from limited image support.
- **An analysis of image-quality metrics for policy evaluation.** We analyze the relationship between reconstruction quality and real-to-sim policy-ranking agreement using standard image-quality metrics (PSNR, SSIM [Wang et al. \[2004\]](#), and LPIPS [Zhang et al. \[2018\]](#)), alongside NVS-SQA.

II. Related Work

Real-to-sim policy evaluation. SIMPLER [Li et al. \[2025\]](#) evaluates whether visually matched simulation preserves policy performance comparisons, using success-rate correlation and mean maximum rank violation (MMRV). PolaRiS [Jain et al. \[2025\]](#) and SimFoundry [Ranawaka et al. \[2026\]](#) extend evaluation to reconstructed scenes, including mesh objects and Gaussian-splat backgrounds. We focus on one specific decision: selecting the better of two policies for the same task and scene, measured as the fraction of

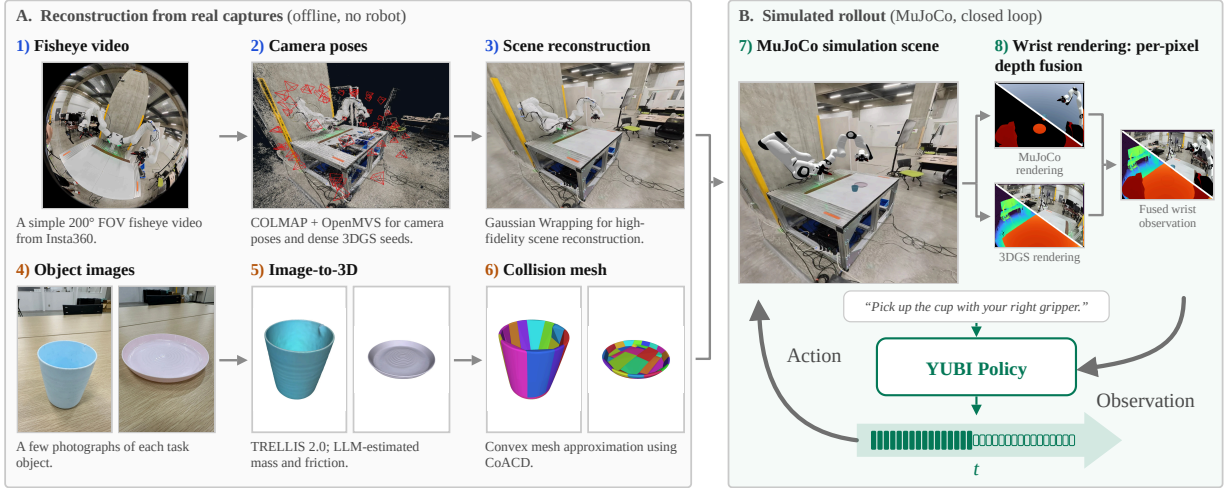


Figure 2 | YUBI-Real2Sim pipeline. We capture the background with a single 200° FOV camera, estimate camera poses with COLMAP, and densify the point cloud with OpenMVS to initialize Gaussian Wrapping for high-fidelity reconstruction. We generate object geometry and textures with TRELIS 2.0 and approximate collision meshes with CoACD. The reconstructed scene and objects are integrated into MuJoCo simulation. For wrist-view rendering, we independently render the background Gaussians and the MuJoCo visual meshes, then composite their pixels using depth maps.

correctly reproduced policy rankings. This pairwise accuracy is distinct from success-rate correlation and MMRV. Related systems reconstruct scenes for policy learning or failure discovery Villasevil et al. [2024]; Han et al. [2025]; Jiang et al. [2025]; Zhang et al. [2025, 2026]; Sagar et al. [2026], while SureSim Badithela et al. [2025] corrects an imperfect simulator using paired real trials. Our experimental variable is the visual reconstruction: we measure whether changing it alters which policy simulation selects.

Evaluating novel-view synthesis. Existing reconstruction-based real-to-sim systems evaluate policy performance in reconstructed scenes Jain et al. [2025]; Ranawaka et al. [2026], but a remaining gap is a validated reconstruction-quality metric for identifying scenes that support reliable policy comparisons. Novel-view synthesis (NVS) provides candidate quality measures, although their relevance depends on the viewpoints being evaluated. Every- N th-frame holdouts mainly measure interpolation; spatial-sector holdouts reveal a 3–12 dB interpolation–extrapolation gap that can reverse reconstruction-method rankings Jia and Appia [2026]. This matters for YUBI, whose independently moving fisheye wrist cameras may observe viewpoints poorly supported by the capture trajectory. When aligned ground-truth views are unavailable, NVS-SQA Qu et al. [2026] scores the perceptual quality of rendered sequences without reference images, while CrossScore Wang et al. [2024] predicts SSIM using unaligned neighboring captures. Fisheye-specific Gaussian Splatting Liao et al. [2024]; Deng et al. [2025]; Huang et al. [2025] addresses wide-field rendering, but rendering quality alone does not establish ranking reliability. We therefore use NVS metrics as candidate diagnostics and test whether their scores identify reconstructed scene variants that fail to preserve real-world pairwise policy rankings.

III. YUBI-Real2Sim System

YUBI-Real2Sim integrates environment capture, static-scene reconstruction, object modeling, and closed-loop policy evaluation under dual fisheye wrist observations (Fig. 2). We project selected fisheye frames into pinhole views, estimate camera poses with COLMAP, generate a dense seed point cloud with OpenMVS, and optimize a 3DGS background with Gaussian Wrapping. Task objects are modeled separately as visual and collision meshes and assembled with the robot in MuJoCo, with the reconstructed background registered to the simulation coordinate system. During simulation, background and mesh renders are composited by depth to form the two calibrated YUBI wrist-camera images. A policy receives these images and proprioception and returns action chunks that drive the simulated robot, closing the loop with the observation modalities and action interface used on hardware. Fig. 2 summarizes the system.

Evaluation setting. We consider three scenes on two bimanual platforms. *Duo* and *Flat* are scenes on a dual Franka FR3 setup and differ in their table and background. The third scene uses a TX-G2 (AgiBot-G2-compatible) platform and is denoted *TX-G2*. Both platforms carry a YUBI gripper with a fisheye camera on each wrist. The task inventory comprises three tasks: Cup (cup pick-and-place), Tape (placing tape into a box), and Pen (placing a pen into a box). The current ranking results use all three tasks; evaluation details appear in Sec. IV.

Capture. Reconstructing the environment for two independently moving fisheye wrist cameras requires visual coverage beyond the immediate manipulation target. Narrow-angle capture can leave surrounding regions poorly observed even though they later appear in policy inputs. We therefore use an Insta360 camera to obtain broad angular coverage with a compact source-frame set, keeping reconstruction optimization practical. Reconstruction uses only its left fisheye lens (Insta360 v6 in the camera metadata). Videos contain two 3840×3840 tracks at 50 fps (HEVC). Capture durations are 213.06 s (*Duo*), 254.88 s (*Flat*), and 243.50 s (*TX-G2*). We partition each video temporally and select one source frame per interval that maximizes the minimum left/right Laplacian variance, without an explicit manual blur-rejection threshold. The same source timestamps also support a controlled comparison between the full fisheye-derived projections and a central pinhole view (Sec. V-C).

Reconstruction. Using OpenCV fisheye calibration, each source frame is reprojected into nine 2048×2048 pinhole views with $90^\circ \times 90^\circ$ FOV and focal length 1024 px: one forward view and eight views tilted 55° from forward, spaced at 45° azimuth intervals. We estimate poses with COLMAP Schönberger and Frahm [2016]; Schönberger et al. [2016] through pycolmap 4.1.1, generate a dense seed point cloud with OpenMVS Cernea [2020], and train Gaussian Wrapping Gomez et al. [2026] at full resolution for 30,000–60,000 iterations using an NVIDIA H200 GPU. The initial dense seed is capped at 40 million points with seed 0. Exposure compensation, normal fields, fast multiview regularization, and masked rendering loss are enabled. The single-lens reconstruction has arbitrary scale. A stored transformation registers it to MuJoCo world coordinates. All embodiments are manually aligned to robot visual meshes, and the tabletop collision surface is modeled as the plane $z = 0$. Collision OBJ meshes are supplied separately from the 3DGS training path. Task objects are reconstructed as separate meshes with TRELLIS 2.0 Xiang et al. [2025] and given CoACD Wei et al. [2022] collision, placed at the real initial layout, and given LLM-estimated mass and friction.

Simulation and rendering. Physics runs in MuJoCo Todorov et al. [2012] in an environment derived from robosuite Zhu et al. [2020], with the robot and objects rendered as meshes and the static background rendered from the 3DGS, composited per pixel by depth. Each YUBI wrist camera is rendered as a cubemap from the 3DGS and remapped to the fisheye image with the intrinsics calibrated on the real YUBI camera (ELP-USB16MP01-L18), at the resolution the policy consumes on the real robot. The camera positions and orientations relative to the wrists are manually fine-tuned in simulation with reference to the physical camera mounts. For each robot configuration, the wrist-relative camera poses are held fixed across the three benchmark tasks, the two candidate policies for each task, and all scene perturbations. This keeps the rendered viewpoint and camera calibration constant, so comparisons across policies and scene variants are not confounded by camera-pose changes. To avoid gradual contact slip inherent in MuJoCo’s soft-contact model Todorov et al. [2012]; MuJoCo, grasps are realized by a contact rule: an object is attached to a gripper when both of its finger inner surfaces touch the object. This attachment is a modeling approximation that bypasses force-dependent grasp stability and suppresses slip-related grasp failures.

Policy interface. $\pi_{0.5}$ Physical Intelligence [2025] policies receive the two fisheye wrist images and proprioception and predict end-effector delta actions referenced to the left hand, together with gripper opening. Actions are predicted in chunks of 32, of which 16 (Cup) or 8 (Tape and Pen) are executed per chunk on both the real robot and in simulation. The FR3 hardware uses impedance control. The TX-G2 setup uses PID control with Ruckig Berscheid and Kröger [2021] on both the real robot and in simulation.

IV. Benchmark Protocol

Scenes and platforms. The benchmark targets three scenes: *Duo* and *Flat* on a dual Franka FR3 setup, and *TX-G2* on a TX-G2 (AgiBot-G2-compatible) platform. Duo and Flat differ in table and background. Both

platforms carry YUBI grippers on both wrists. The same policy runs on both through platform-specific action mappings and low-level controllers.

Tasks. The benchmark contains three tasks: Cup (cup pick-and-place), Tape (placing tape into a box), and Pen (placing a pen into a box). These are the only tasks included in the ranking analysis. A trial is counted as successful as soon as the task-specific success criterion is met within the time limit; otherwise, it is counted as a failure. The time limit is 3 min (1,800 frames) for Tape and Pen in Duo and for Pen in TX-G2, and 1 min (600 frames) for all other scene–task pairs, with actions executed at 10 Hz. Within each scene–task pair, the same time limit applies to both policies in real evaluation and across all simulated reconstruction conditions. For real trials, an operator assigns the label using a fixed task-level rubric; for simulation, a deterministic object-state predicate implements the corresponding endpoint. The rubric and simulation predicate are held fixed across policies within each task.

Policies and counting unit. Each task has two task-specific $\pi_{0.5}$ -based candidates, denoted π_1 and π_2 with labels fixed before comparison. They are not two shared checkpoints across tasks. Three tasks on three scenes provide 18 policy cells and nine potential pairwise comparisons, all of which are evaluated in the current result.

Trials. Each policy cell uses 20 real and 80 simulated rollouts. Within each scene–task comparison, both policies and the clean and perturbed variants share the initial layout and initial-condition/seed schedule. Objects are reset to the same nominal layout before each trial. Object assets and the simulated layout are matched to the corresponding real setup. Simulated rollouts use different seeds for VLA inference, with the same seed list reused across clean and perturbed scenes as specified in Sec. IV-A.

IV-A. Evaluation conditions

The evaluation has a clean reconstruction condition and two types of counterfactual condition. The artifact conditions edit the reconstructed background while holding the robot, physics, collision geometry, object assets, initial conditions, policy checkpoints, success rules, and rollout seeds fixed. The reduced-coverage conditions retrain the 3DGS scene with fewer source images and are used for visual defect, quality, and policy-ranking analyses.

Clean and artifact conditions. The clean reconstruction uses all nine projected views per selected left-lens source frame. The artifact design contains six variants per scene: two holes, three floater configurations, and one warping realization. Each edit changes only background Gaussians, leaving collision geometry and object assets fixed. Initial conditions, policy checkpoints, success rules, and rollout seed lists are paired with the clean evaluation.

Hole: an upward cone removes Gaussians at half-angles of 54° and 90° from the vertical. The table protection region uses its collision footprint expanded by 40 mm, a ± 60 mm slab about its surface, and below-table protection; the ceiling directly above the table remains eligible for removal.

Floater: eight clusters of four Gaussians with neighboring scene colors are placed over the table in each of three height bands: 5–15, 15–25, and 25–35 cm. Placement maintains 10 mm clearance from the table and objects. Within-cluster positional standard deviation is 18 mm, the Gaussian size center is 31.6 mm, and the opacity center is 0.99; view-dependent SH is disabled. These are three artifact configurations; isolating height alone additionally requires matching their horizontal positions, sizes, and colors.

Warping: one fixed seed places 30 smooth domes that displace table-region Gaussians vertically, with sampled heights of 10–30 mm and radii of 30–60 mm. The region uses the footprint expanded by 40 mm and a ± 60 mm surface band. Overlapping displacements add without a cap, so the total displacement may exceed 30 mm. One seed characterizes one fixed realization rather than a distribution over random table defects.

Reduced-coverage conditions. The supporting reconstruction study compares Clean, Central pinhole, Contiguous 30% removal, and Random 30% removal reconstructions at the same held-out and robot-view protocols. Central pinhole retains only the forward left-lens projection (L_f) at every source timestamp. Both removal conditions also use only L_f : Contiguous 30% removal drops one consecutive interval containing 30% of the time-ordered selected source frames, and Random 30% removal drops the same number at random. The interval location and random seed are fixed before examining reconstructed artifacts.

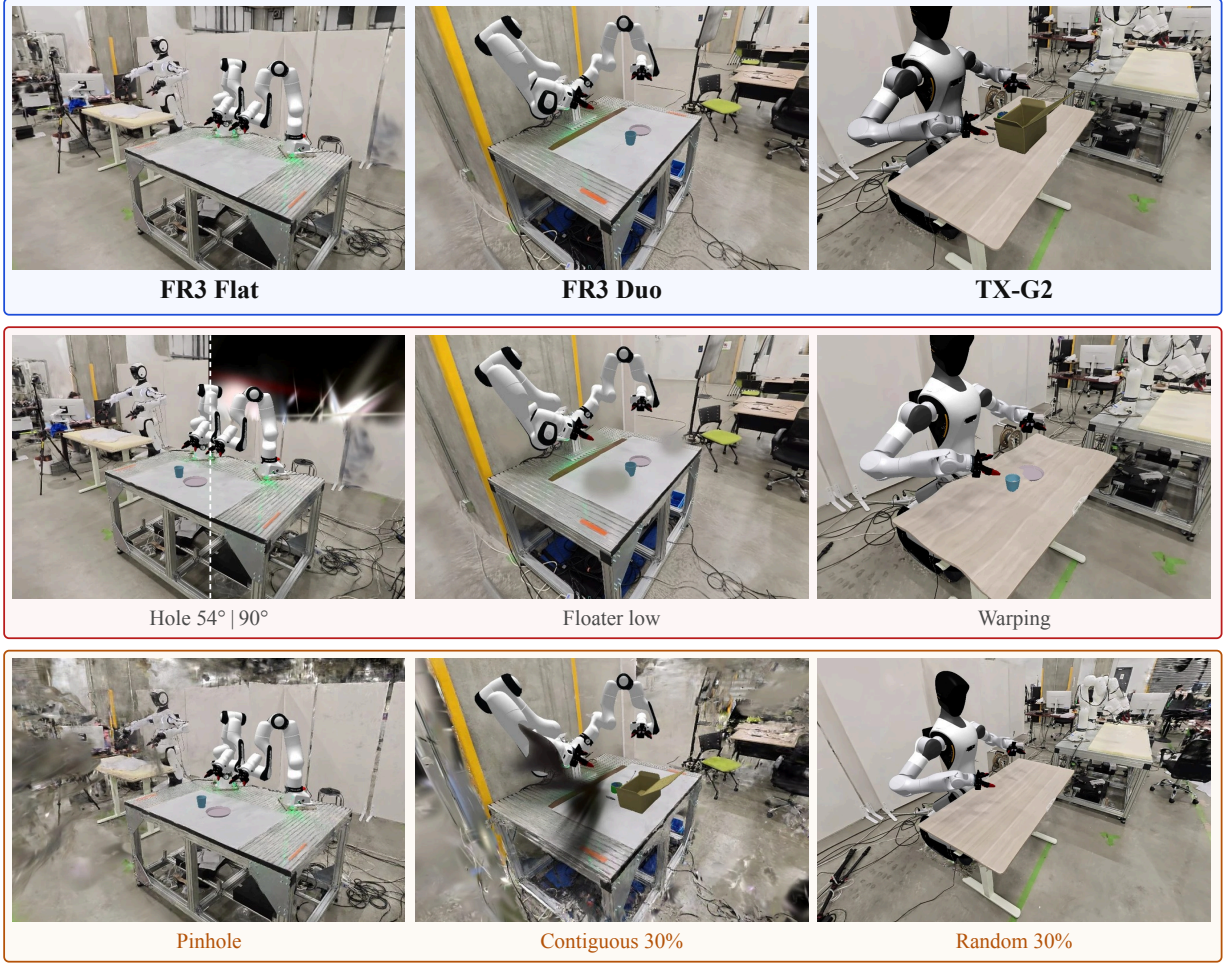


Figure 3 | Simulated scenes, and the added disturbances for 3DGS or viewpoints. One column per scene, one band per condition. **Blue:** the clean 3DGS reconstructions the rollouts run in. **Red:** the same reconstructions carrying a counterfactual artifact — a ceiling hole in FR3 Flat, the 54° and 90° half-angles either side of the dashed rule, low floaters in FR3 Duo, and warping in TX-G2. **Amber,** the colour Fig. 1 gives capture: one reduced-coverage reconstruction per condition, each in the scene of its column. The panels show representative examples of the reconstruction conditions. We evaluated two task-specific VLA checkpoints for each of three tasks (Cup, Tape, and Pen) in all three scenes (FR3 Flat, FR3 Duo, and TX-G2).

For a scene with N source frames, the removal conditions use

$$n_{\text{remove}} = \text{round}(0.30N), \quad N_{\text{keep}} = N - n_{\text{remove}}. \quad (1)$$

Removal is applied to the central-pinhole source-frame set. This leaves 134/192 frames for Duo, 105/150 for Flat, and 153/219 for TX-G2. These counts precede splitting; quality-evaluation views are held fixed and excluded from every training set. The central-view condition changes projections per source frame from nine to one; each projection remains 2048×2048 with a $90^\circ \times 90^\circ$ FOV.

COLMAP intrinsics, registered poses, and the original dense seed are reused. Each condition uses fresh optimization with seed 0 and 30,000 iterations. The same held-out views are used across all reconstruction conditions and excluded from training. Clean, edited, and reduced-coverage reconstructions are compared at the same camera poses, including wrist views, and visible defect type, location, and extent are documented. Visual similarity alone does not establish a policy-ranking effect for a retrained variant.

IV-B. Metrics and reporting

For scene s , task t , policy $p \in \{\pi_1, \pi_2\}$, and domain $d \in \{\text{real}, \text{sim}\}$, the success-rate estimate is

$$\hat{q}_{s,t,p}^d = \frac{S_{s,t,p}^d}{N_{s,t,p}^d}. \quad (2)$$

Pairwise ranking accuracy. For scene-task pair i , let r_{i,π_1}, r_{i,π_2} and s_{i,π_1}, s_{i,π_2} be observed real and simulated success rates. Define

$$d_i^{\text{real}} = r_{i,\pi_1} - r_{i,\pi_2}, \quad d_i^{\text{sim}} = s_{i,\pi_1} - s_{i,\pi_2}. \quad (3)$$

Let M contain pairs with both required evaluations and a strict observed real ranking ($d_i^{\text{real}} \neq 0$). Our primary metric is

$$\text{Acc}_{\text{rank}} = \frac{1}{|M|} \sum_{i \in M} \mathbf{1}[d_i^{\text{real}} d_i^{\text{sim}} > 0]. \quad (4)$$

The numerator is the number k of eligible comparisons for which simulation selects the same policy as the real evaluation, so $\text{Acc}_{\text{rank}} = k/|M|$. Thus this metric measures agreement in the two-policy choice, not agreement in absolute success rates. A reversed order and a simulation tie both score zero; we report them separately. Real ties are unrankable and are reported separately. The comparison set is fixed before testing scene perturbations, so a variant with two zero-success policies remains in the denominator as a simulation tie once that comparison is completed. We report correct/total pairs per scene and overall, weighting each eligible scene-task pair equally. Success counts support the ranking; matching their absolute values is outside this benchmark’s claim.

The success-rate bars show marginal 95% Wilson score intervals using 20 real or 80 simulated trials per policy. These intervals quantify each success-rate estimate; they are not confidence intervals for ranking accuracy or tests of the difference between policies. Ranking accuracy is reported as a point estimate over the fixed set of eligible scene-task comparisons; rollout counts are not treated as independent ranking comparisons.

We also report the absolute-rate error,

$$\text{MAE} = \frac{1}{2|M|} \sum_{i \in M} \sum_{p \in \{\pi_1, \pi_2\}} |\hat{q}_{i,p}^{\text{real}} - \hat{q}_{i,p}^{\text{sim}}|. \quad (5)$$

This secondary metric does not determine whether the policy ranking is correct. Only the Clean MAE is reported (Sec. V-A).

Counterfactual comparison. For artifact variant v , we report its ranking-accuracy change from clean as

$$\Delta \text{Acc}_{\text{rank}}(v) = \text{Acc}_{\text{rank}}(v) - \text{Acc}_{\text{rank}}(\text{clean}). \quad (6)$$

We additionally count clean-correct comparisons that become reversed or tied, and clean-incorrect comparisons that become correct. A decrease in both policies’ success rates is a rate change only; it is a ranking failure when the policy ranking changes or becomes tied. Table I reports these counts over the nine comparisons of each condition.

Quality metrics and view protocols. For metric m , scene s , and view protocol u , we report the paired score change

$$\Delta Q_{m,s,u} = Q_{m,s,u}^{\text{counterfactual}} - Q_{m,s,u}^{\text{clean}}. \quad (7)$$

NVS-SQA Qu et al. [2026] is evaluated on held-out Test views and OBS views. Test views have aligned

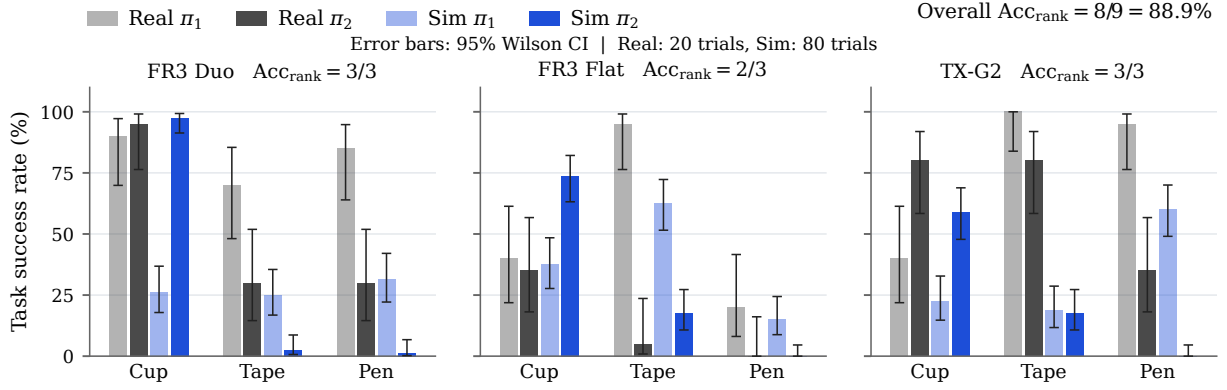


Figure 4 | Policy ranking in real and simulated evaluation. Black/gray bars show real success and blue bars show simulated success; light and dark bars denote π_1 and π_2 , respectively, as in Fig. 1. Policy labels are fixed within each task across scenes and domains. Error bars are marginal 95% Wilson confidence intervals from 20 real and 80 simulated trials per policy. The observed ranking agrees in 8/9 scene-task pairs (88.9%); Flat Cup is reversed.

captured images, so PSNR, SSIM Wang et al. [2004], and LPIPS Zhang et al. [2018] are also reported there; OBS has no paired ground-truth image. OBS-Probe uses a fixed pre-rollout trajectory, whereas OBS-Policy uses frames observed during policy rollouts. Scores are averaged over frames within each scene or trajectory.

Scene verification without target-scene real evaluation. For task t in scene variant (s, v) , $H_{s,t,v} = 1$ denotes disagreement between the simulated and observed real policy ranking, including a simulation tie. We predict this label from the available scene-level quality score:

$$\hat{H}_{s,t,v} = \mathbf{1}[Q_{s,v} < \tau_{-s,t}]. \quad (8)$$

A separate threshold for each task is selected on the other two scenes by maximizing balanced accuracy, with sensitivity and then the smallest threshold as tie-breakers. All variants of the target scene are held out together. If training labels contain only one class, the predictor returns that class; no quality threshold is learned in that fold. Target-scene real outcomes are used only to score predictions. Test and OBS use identical eligible conditions within each comparison. Test PSNR, SSIM, and LPIPS use the same threshold selection and held-out conditions; LPIPS is negated so that higher scores consistently indicate better quality. These three metrics use aligned captured reference images for Test views. A second scene-level label requires all three task rankings to agree with real evaluation. The available OBS scores are shared scene-variant measurements, rather than separate task-specific scores.

V. Policy Ranking and Reconstruction Quality

Evaluation questions. The experiments answer four separate questions. The clean experiment asks whether simulation selects the same policy as the real robot. The injected artifact experiment asks whether reconstruction defects change that selection. The view-count experiment asks whether reducing capture support produces similar visual defects and, when rollouts are run, changes policy selection. The quality experiment asks whether image-quality scores can identify scenes whose policy rankings are unreliable. Each question has a separate outcome; a lower task success rate alone does not imply a ranking failure.

V-A. Clean policy ranking

Q: Can clean simulation reproduce which of π_1 and π_2 performs better on the real robot?

A: Yes. It reproduces the real policy ranking in 8 of 9 scene-task comparisons (88.9%), with one inversion in Flat/Cup. The success-rate estimate, ranking metric, and secondary absolute-rate error are defined in the protocol (Eqs. 2-5). Simulation ties and reversals are counted as failures of the choice and are reported separately.

Table I | Effect of each counterfactual (CF) condition on the nine scene–task policy comparisons. Significance: comparisons whose Fisher-test status ($p < 0.05$) is lost or gained relative to Clean; the others keep their status. Ranking vs. Clean: comparisons whose simulated ranking is reversed or tied relative to Clean. Agreement with real: comparisons whose simulated ranking matches the real ranking. Clean agrees in 8/9 (Flat Cup is reversed) and no CF condition corrects it, so every reversal or tie is a lost agreement.

CF condition	Significance		Ranking vs. Clean		Agreement with real
	Lost	Gained	Reversed	Tied	
Hole 54°	1	1	1	0	7/9
Hole 90°	4	0	1	0	7/9
Floater low8	1	0	0	1	7/9
Floater mid8	1	1	1	0	7/9
Floater high8	1	0	1	0	7/9
Warping	2	1	0	0	8/9
Central pinhole	5	0	0	0	8/9
Contiguous 30% removal	3	0	1	0	7/9
Random 30% removal	3	0	2	0	6/9

Fig. 4 compares real and simulated success rates for the two policies on Cup, Tape, and Pen, and summarizes their rank agreement. Eight of nine scene–task pairs agree with real evaluation, giving 88.9% pairwise ranking accuracy. The single inversion is Flat Cup: real success is 40% versus 35%, while simulated success is 37.5% versus 73.75% at 600 frames for the same policy labels. Clean MAE is 31.3889 percentage points (0.3139 on the unit-interval scale); absolute-rate error and pairwise ranking agreement are distinct outcomes, and neither substitutes for the other.

V-B. Counterfactual artifacts

Q: How do injected reconstruction artifacts change policy-ranking agreement and the significance of policy success-rate differences?

A: Holes and floaters reduce agreement from 8/9 to 7/9; Warping preserves 8/9 agreement, although significance status changes. Table I evaluates all nine scene–task comparisons for each condition, with 80 simulated trials per policy. Transitions use the Clean reference results in Sec. V-A as the common reference. Both Hole angles and all three Floater configurations yield 7/9 (77.8%) ranking accuracy, compared with Clean’s 8/9 (88.9%). The additional disagreement is TX-G2 Tape: Hole 54°, Hole 90°, and Floater mid8/high8 reverse its ranking, while Floater low8 makes the policies tie. Warping preserves all Clean rankings and hence 8/9 accuracy.

Ranking agreement and statistical significance capture different outcomes. At the two-sided Fisher-test threshold $p < 0.05$, Hole 90° removes significance in four comparisons: Duo Tape, Duo Pen, Flat Pen, and TX-G2 Pen. For example, Duo Pen changes from 25/80 versus 1/80 in Clean ($p < 0.001$) to 7/80 versus 4/80 ($p = 0.534$), retaining its ranking. Conversely, Hole 54° reverses TX-G2 Tape from 15/80 versus 14/80 ($p = 1$) to 1/80 versus 9/80 ($p = 0.018$), producing a significant ranking opposite to the observed real ranking. Warping retains the TX-G2 Tape ranking while making the difference significant (14/80 versus 5/80, $p = 0.048$). Crossing the significance threshold is not itself a test that the policy success-rate gap has changed significantly.

V-C. Reconstruction from reduced image coverage

Q: How does reduced image coverage affect policy-ranking agreement?

A: Central pinhole retains 8/9 agreement, while Contiguous 30% removal and Random 30% removal reduce it to 7/9 and 6/9, respectively. As summarized in Table I, Contiguous 30% removal reverses TX-G2 Cup; Random 30% removal reverses both TX-G2 Cup and TX-G2 Tape. Random 30% removal therefore has the largest observed accuracy decrease, from 88.9% to 66.7% (22.2 percentage points). Central pinhole preserves all Clean rankings but removes policy success-rate significance in five comparisons, illustrating that unchanged ranking accuracy can coexist with weaker statistical separation. Contiguous and Random 30% removal each remove significance in three comparisons. All reduced-coverage conditions use the same per-task time limits as Clean (Sec. IV).

Table II | Scene-held-out reliability classification: correct/total decisions on 15 matched conditions per task. NVS: NVS-SQA; Total: pooled task decisions; All-three: reliability of all three task rankings.

Metric/view	Cup	Tape	Pen	Total	All-three
NVS Test	7/15	8/15	15/15	30/45	13/15
NVS OBS	12/15	2/15	15/15	29/45	7/15
PSNR Test	8/15	8/15	15/15	31/45	9/15
SSIM Test	8/15	9/15	15/15	32/45	7/15
LPIPS Test	8/15	5/15	15/15	28/45	10/15

V-D. Quality response to view coverage

Q: Does NVS-SQA reveal how each view-coverage disturbance changes scene quality relative to clean reconstruction?

A: Weakly. The change is scene- and mode-dependent and does not share a sign, but it is small against the spread between scenes.

On held-out Test views, Contiguous 30% removal moves mean NVS-SQA by -0.060 for Duo, $+0.031$ for Flat, and -0.018 for TX-G2; Random 30% removal by $+0.028$, $+0.009$, and -0.039 ; Central pinhole by $+0.018$, -0.005 , and $+0.004$. Every one of these is far smaller than the 0.224 spread between the three scenes’ own clean scores, so the metric separates scenes more strongly than it separates a scene’s capture conditions. The score-change definition and the Test/OBS view protocols are specified in the benchmark protocol.

V-E. Scene verification across tasks and scenes

Q: Can scene-quality scores identify reliable policy rankings across Cup, Tape, and Pen without target-scene real evaluation?

A: OBS improves Cup decisions, but this advantage does not extend to Tape or to the aggregate current evaluation. We evaluate Cup, Tape, and Pen across three scenes and five matched variants: Hole $54^\circ/90^\circ$, Central pinhole, Contiguous 30% removal, and Random 30% removal. We compare NVS-SQA on Test/OBS with PSNR, SSIM, and LPIPS on Test, selecting task-specific thresholds on two scenes and evaluating the third. Table II reports reliability-classification accuracy. Among the five quality-based methods in Table II, SSIM achieves the highest pooled task-wise accuracy (32/45), while Test NVS-SQA performs best for deciding whether all three task rankings are reliable (13/15). NVS-SQA on OBS improves Cup decisions but not aggregate accuracy (29/45 versus 30/45 for Test). Pen’s perfect score reflects constant predictions because all its labels are reliable, rather than demonstrated failure detection. The always-reliable predictor achieves 35/45 on the same conditions, exceeding all five quality-based methods in pooled accuracy. All three Tape disagreements occur in TX-G2, so the TX-G2 fold trains on a single class and every quality-based method misses them; Tape accuracy differences therefore arise only from false alarms on agreement-labeled conditions.

VI. Conclusion

We presented YUBI-Real2Sim to evaluate whether reconstructed scenes support reliable policy evaluation without real-robot verification in the target scene. Clean simulation agrees with real policy rankings in 8/9 scene-task comparisons. Agreement is 7/9 under holes and floaters and 6/9 under Random 30% removal; unchanged rankings can nevertheless lose statistical separation between policies. Scene-held-out evaluation uses quality scores to predict reliability from real labels in other scenes. OBS-based NVS-SQA improves Cup decisions, but the benefit does not extend to the aggregate three-task result. The results distinguish agreement with real rankings, statistical separation of policies, and quality-based prediction of reliability as separate evaluation outcomes.

Acknowledgment

This paper is based on results obtained from a project, JPNP25015, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

- Apurva Badithela, David Snyder, Lihan Zha, Joseph Mikhail, Matthew O’Kelly, Anushri Dixit, and Anirudha Majumdar. Reliable and scalable robot policy evaluation with imperfect simulators. *arXiv preprint arXiv:2510.04354*, 2025.
- Lars Berscheid and Torsten Kröger. Jerk-limited real-time trajectory generation with arbitrary target states. In *Proceedings of Robotics: Science and Systems*, July 2021. doi: 10.15607/RSS.2021.XVII.015. URL <https://www.roboticsproceedings.org/rss17/p015.html>.
- Dan Cernea. OpenMVS: Multi-view stereo reconstruction library. <https://github.com/cdcseacave/openMVS>, 2020. Accessed: 20 September 2026.
- Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. In *RSS*, 2024.
- Youming Deng, Wenqi Xian, Guandao Yang, Leonidas Guibas, Gordon Wetzstein, Steve Marschner, and Paul Debevec. Self-calibrating gaussian splatting for large field of view reconstruction. *arXiv preprint arXiv:2502.09563*, 2025.
- Diego Gomez, Antoine Guédon, Nissim Maruani, Bingchen Gong, and Maks Ovsjanikov. From blobs to spokes: High-fidelity surface reconstruction via oriented gaussians. *arXiv preprint arXiv:2604.07337*, 2026.
- Xiaoshen Han, Minghuan Liu, Yilun Chen, Junqiu Yu, Xiaoyang Lyu, Yang Tian, Bolun Wang, Weinan Zhang, and Jiangmiao Pang. Re³sim: Generating high-fidelity simulation data via 3d-photorealistic real-to-sim for robotic manipulation. *arXiv preprint arXiv:2502.08645*, 2025.
- Huajian Huang, Yingshu Chen, Longwei Li, Hui Cheng, Tristan Braud, Yajie Zhao, and Sai-Kit Yeung. SC-OmniGS: Self-calibrating omnidirectional gaussian splatting. *arXiv preprint arXiv:2502.04734*, 2025.
- 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.
- Gaoxiang Jia and Vikram Appia. Mind the gap: Standard 3DGS evaluation primarily measures near-trajectory interpolation. *arXiv preprint arXiv:2607.01556*, 2026. doi: 10.48550/arXiv.2607.01556. URL <https://arxiv.org/abs/2607.01556>.
- Guangqi Jiang, Haoran Chang, Ri-Zhao Qiu, Yutong Liang, Mazeyu Ji, Jiyue Zhu, Zhao Dong, Xueyan Zou, and Xiaolong Wang. GSWorld: Closed-loop photo-realistic simulation suite for robotic manipulation. *arXiv preprint arXiv:2510.20813*, 2025.
- Shi Jin, Yuntian Wang, Yuhui Duan, Di Wu, et al. UMI-Bench 1.0: An open and reproducible real-world benchmark for tabletop robotic manipulation with UMI data. *arXiv preprint arXiv:2606.10382*, 2026.
- Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023.
- Xuanlin Li, Kyle Hsu, Jiayuan Gu, Oier Mees, Karl Pertsch, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating real-world robot manipulation policies in simulation. In *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>.

- Zimu Liao, Siyan Chen, Rong Fu, Yi Wang, Zhongling Su, Hao Luo, Li Ma, Linning Xu, Bo Dai, Hengjie Li, Zhilin Pei, and Xingcheng Zhang. Fisheye-GS: Lightweight and extensible gaussian splatting module for fisheye cameras. *arXiv preprint arXiv:2409.04751*, 2024.
- MuJoCo. MuJoCo documentation: Modeling—preventing slip. Online documentation. URL <https://mujoco.readthedocs.io/en/stable/modeling.html#preventing-slip>. Accessed: 2026-09-25.
- Takehiko Ohkawa, Jumpei Arima, Yuki Noguchi, Masatoshi Tatenno, Makoto Sugiura, Takuya Okubo, Kengo Ikeuchi, Yuma Shin, Hiroki Nishizawa, Naoaki Kanazawa, Yuki Wakayama, Daiki Fukunaga, Koshi Makihara, Tomohiro Motoda, Floris Erich, Yukiyasu Domae, Tatsuya Matsushima, Yohishiro Okumatsu, and Kei Ota. YUBI: Yielding universal bidigital interface for bimanual dexterous manipulation at scale. *arXiv preprint arXiv:2606.10244*, 2026.
- Physical Intelligence. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.
- Qiang Qu, Yiran Shen, Xiaoming Chen, Yuk Ying Chung, Weidong Cai, and Tongliang Liu. NVS-SQA: Exploring self-supervised quality representation learning for neurally synthesized scenes without references. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):2265–2281, 2026.
- Nadun Ranawaka, Josiah Wong, Wei-Lin Pai, Wei-Teng Chu, Tianyuan Dai, Masoud Moghani, Hang Yin, Yunfan Jiang, Wesley Durbano, Brandon Huynh, Yu Fang, Linxi Fan, Danfei Xu, Ruohan Zhang, Li Fei-Fei, Bowen Wen, Ajay Mandlekar, and Yuke Zhu. SimFoundry: Modular and automated scene generation for policy learning and evaluation. *arXiv preprint arXiv:2606.28276*, 2026.
- Som Sagar, Jiafei Duan, Sreevishakh Vasudevan, Yifan Zhou, Heni Ben Amor, Dieter Fox, and Ransalu Senanayake. RoboMD: Uncovering robot vulnerabilities through semantic potential fields. *arXiv preprint arXiv:2412.02818v4*, 2026.
- Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016.
- Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *IROS*, pages 5026–5033, 2012.
- Marcel Torne Villasevil, 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 *RSS*, 2024.
- Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004. URL <https://ece.uwaterloo.ca/~z70wang/research/ssim/>.
- Zirui Wang, Wenjing Bian, and Victor Adrian Prisacariu. CrossScore: Towards multi-view image evaluation and scoring. In *ECCV*, 2024.
- Xinyue Wei, Minghua Liu, Zhan Ling, and Hao Su. Approximate convex decomposition for 3D meshes with collision-aware concavity and tree search. *ACM Transactions on Graphics*, 41(4):1–18, 2022. doi: 10.1145/3528223.3530103. URL <https://doi.org/10.1145/3528223.3530103>.
- Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, and Jiaolong Yang. Native and compact structured latents for 3D generation. *arXiv preprint arXiv:2512.14692*, 2025. doi: 10.48550/arXiv.2512.14692. URL <https://arxiv.org/abs/2512.14692>.
- Kaifeng Zhang, Shuo Sha, Hanxiao Jiang, Matthew Loper, Hyunjong Song, Guangyan Cai, Zhuo Xu, Xiaochen Hu, Changxi Zheng, and Yunzhu Li. Real-to-sim robot policy evaluation with gaussian splatting simulation of soft-body interactions. *arXiv preprint arXiv:2511.04665*, 2025.
- Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on*

Computer Vision and Pattern Recognition (CVPR), 2018. URL <https://richzhang.github.io/PerceptualSimilarity/>.

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.

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.