RFW-W: SCALING PHYSICAL ASSET UNIVERSES FOR GENERATIVE EMBODIED SIMULATION

Anonymous authors
Paper under double-blind review



Figure 1: **RFW-W: from physical assets to embodied tasks.** Left: RFW-R provides diverse rigid assets with physical and interaction annotations. Center: task-conditioned retrieval and composition organize these assets into simulation scenes. Right: paired real and simulated setups illustrate shared manipulation objectives and object relations. Together, these components connect asset coverage and executable representations to scene construction and robot manipulation.

ABSTRACT

Broad object coverage and executable physical representations are essential for scaling embodied simulation. Generative 3D models expand visual diversity, but scene construction and manipulation also require physical and interaction annotations. We present RFW-W, which combines RFW-R, a large-scale rigid-object dataset, with a common physical-asset interface and task-conditioned scene composition. RFW-R contains 524,611 retained assets across 1,212 categories, with collision geometry, composed physical properties, and interaction annotations exposed through standardized simulation interfaces. For each task, the system retrieves semantically relevant, physically compatible assets and assembles them around the specified object relations. For scene construction on 16 shared room layouts, RFW-R supplies and places 100% of the 265 requested objects, compared with 97.4% for EmbodiedGen v2 and 90.6% for ManiTwin. Across five RoboTwin tasks with 100 demonstrations per task and training source, RFW-R-trained policies achieve 71.2% clean and 41.8% randomized macro success, exceeding the strongest baseline by 15.8 and 10.6 percentage points, respectively. These results demonstrate the utility of RFW-R for scene construction and simulated policy learning.

1 INTRODUCTION

General-purpose embodied intelligence requires robots to operate across diverse objects, layouts, and interactions. Physics simulation provides a scalable alternative to real-world data collection, enabling reusable and varied environments for embodied learning (Kolve et al., 2017; Deitke et al., 2022; Nasiriany et al., 2024; Wang et al., 2023b; Xia et al., 2026; Wang et al., 2026c). Yet its effective scope depends on the underlying asset library: scene generation and manipulation require not only diverse objects, but also representations that physics simulators can reliably execute.

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.

Table 1: **Comparison of rigid asset libraries.** *Part phys.*: per-part mass, density, and friction released as numerical fields, with inertia and center of mass composed and verified; *Hollow*: hollow-aware mass modeling; *Med. inst.*: median instances per class; *Scene fill*: instance fulfillment over 265 frozen requests in 16 foundations (Section 4.5); *Grasp ann.*: released grasp annotations with simulation checks. RFW-R is the rigid collection constructed in this work and combines part-level physicalization, scene-scale coverage, and dexterous-hand grasp annotations.

Dataset	Scene-scale coverage				Part-level physics			Dexterous manip.	
	#Assets	#Classes	Med. inst.	Scene fill	Seg.	Part phys.	Hollow	Hands	Grasp ann.
EmbodiedGen v2 (Wang et al., 2026c)	4,168	200	11	97.4%	✓	placeholder	×	0	×
PhysX-3D (Cao et al., 2025a)	32,047	450	1	–	✓	density	×	0	×
ManiTwin (Wang et al., 2026b)	4,574	485	6	90.6%	×	×	×	0	✓
RFW-R (this work)	**524,611**	**1,212**	**356**	**100%**	✓	✓	✓	**6**	✓

Large-scale 3D repositories and generative models provide millions of assets and open-category synthesis of geometry and appearance (Chang et al., 2015; Deitke et al., 2023b;a; Lin et al., 2025; Zhang et al., 2024b; Xiang et al., 2024; 2025; Li et al., 2025; Team Hunyuan3D et al., 2025). However, visual meshes alone do not specify metric scale, collision geometry, mass, friction, inertia, or interaction semantics. Collections such as YCB and Google Scanned Objects provide physical object models for manipulation and simulation (Calli et al., 2015; Downs et al., 2022), but extending these representations across broad categories and hundreds of thousands of assets remains challenging.

Recent physical-asset datasets and articulation-modeling methods enrich 3D content with physical, functional, and articulation information (Cao et al., 2025a;b; 2026; Yang et al., 2026b; Li et al., 2026; Iliash et al., 2026; Yang et al., 2026a). ManiTwin adds manipulation-oriented digital twins, while Seed3D and EmbodiedGen connect assets to broader simulation-generation systems (Wang et al., 2026b; Gu et al., 2026; Wang et al., 2026c). Nevertheless, existing libraries exhibit trade-offs among category coverage, part-level physicalization, and dexterous manipulation readiness (Table 1), with category distributions often inherited from their source repositories. We therefore ask: *can these capabilities be jointly scaled across a broad rigid-asset taxonomy, and does the resulting collection improve scene construction and manipulation policy learning?*

To study this problem at scale, we introduce **RoboFlywheel (RFW-W)**: a common physical-asset interface, a large-scale rigid collection, and a task-conditioned system for retrieval and scene composition. We construct **RFW-R**, a rigid-object dataset, under an explicit, extensible category taxonomy, retaining 524,611 rigid assets across 1,212 leaf categories with a median of 356 instances per category. Each asset supplies part semantics, collision geometry, metric scale, material and friction estimates, mass, center of mass, inertia, and interaction annotations. RFW-R also provides hollow-aware physical modeling, simulator export, cross-simulator checks, and simulation-validated grasps for six dexterous-hand configurations (Table 1). RFW-W exposes RFW-R assets through a common asset schema for retrieval and scene composition.

We evaluate whether the breadth and physical executability of RFW-R translate into downstream utility. We measure asset and part coverage, fixed-request scene fulfillment, simulated grasp execution, and policy learning. RFW-R fulfills all 265 requested instances across 16 fixed scene foundations (Table 1); RFW-R-trained policies also lead the matched-demonstration comparison across five RoboTwin tasks (Table 5). These experiments test the rigid collection and the scene-construction system. The reported inventory and asset-level gains concern RFW-R and its evaluated protocols.

Our contributions are summarized as follows:

- We construct a verifiable physical asset lifecycle that transforms open-category rigid assets into part-level, collision-ready, inertially consistent, and cross-simulator executable representations.

- We bridge rigid assets and embodied worlds through a common physical-asset interface and an agentic generation system that composes RFW-R assets into task-conditioned simulation environments.
- We move beyond instance-count scaling and formulate category breadth as a first-class scaling variable, demonstrating its importance for scene coverage, manipulation generalization, and sim-to-real transfer.

2 Related Work

2.1 From 3D Assets to Physical Asset Universes

Large-scale 3D repositories have substantially expanded the scale, category coverage, and visual diversity of available 3D content Chang et al. (2015); Fu et al. (2020); Collins et al. (2021); Wu et al. (2023); Deitke et al. (2023b;a); Lin et al. (2025), while recent generative models further enable open-category synthesis of high-quality geometry, textures, and PBR materials Zhang et al. (2024b); Xiang et al. (2024; 2025); Team Hunyuan3D et al. (2025); Li et al. (2025). Building on these advances, recent physical-asset datasets augment 3D objects with metric scale, materials, physical properties, affordances, functional semantics, and simulation-ready representations (Cao et al., 2025a;b; Yang et al., 2026b; Li et al., 2026). ManiTwin further provides manipulation-oriented annotations and verified interaction proposals, while Seed3D and EmbodiedGen connect physical assets to broader embodied-generation systems (Wang et al., 2026b; Feng et al., 2025; Gu et al., 2026; Wang et al., 2025; 2026c). Despite this progress, existing collections still exhibit trade-offs among category breadth, long-tail coverage, and intra-category diversity. RFW-R complements these efforts by scaling a rigid physical-asset collection with explicit and controllable category coverage.

Earlier manipulation collections establish useful precedents: YCB associates physical objects with geometric and physical models (Calli et al., 2015), while Google Scanned Objects prepares scanned household items for simulation. RFW-R extends this asset-centered line toward broader rigid-category and instance coverage with part-level interaction representations.

2.2 Executable Asset Representations for Embodied Interaction

Physical executability requires more than a visual mesh. PartNet, SAPIEN, and GAPartNet establish structured and actionable part representations. Simulation-ready assets additionally require dedicated collision geometry and consistent physical properties such as scale, mass, center of mass, inertia, and friction (Mo et al., 2019; Xiang et al., 2020; Geng et al., 2023; Wei et al., 2022). Complementary work provides large-scale grasp and interaction annotations, from ACRONYM and DexGraspNet to GraspXL, ManiTwin, and EmbodiedGen (Eppner et al., 2020; Wang et al., 2022; Zhang et al., 2024a; Wang et al., 2026b;c). RFW-R combines these structural, physical, and interaction components with standardized simulator interfaces and execution-based validation at scale.

2.3 From Physical Assets to Generative Embodied Worlds

Procedural and language-driven scene generation has enabled increasingly diverse and scalable embodied environments through structured generation, retrieval, learned layouts, and physical constraints (Deitke et al., 2022; Raistrick et al., 2023; 2024; Feng et al., 2023; Yang et al., 2023; Fu et al., 2023; Yang et al., 2024; Lin & Mu, 2024). More recent agentic systems introduce iterative planning, tool use, and visual or physical feedback to construct task-conditioned, simulation-ready worlds (Ling et al., 2025; Yang et al., 2025b; Zhao et al., 2026; Xia et al., 2026; Wang et al., 2026c). In parallel, simulation platforms and data engines scale environments, tasks, and demonstrations for robot learning (Szot et al., 2021; Puig et al., 2023; Li et al., 2021; 2023; Gu et al., 2023; Tao et al., 2024; Nasiriany et al., 2024; Chen et al., 2025; Mandlekar et al., 2023; Wang et al., 2023a;b). While these efforts primarily scale downstream worlds and interaction data, RFW-W studies how an upstream rigid-asset collection and shared retrieval interface support scene construction and manipulation policy learning.

AI2-THOR provides interactive indoor environments (Kolve et al., 2017), and 3D-FRONT provides a large collection of furnished indoor layouts (Fu et al., 2021). TableVerse reconstructs simulation-ready tabletop scenes from real-world images, instantiates assets, checks stability, and generates

manipulation trajectories (Wang et al., 2026a). These resources make scene structure and object arrangement reusable; our scene pipeline additionally tests whether retrieved objects satisfy task requests and carry executable physical properties.

2.4 Robot Policies and Training Data

Diffusion Policy models visuomotor actions with a conditional diffusion process (Chi et al., 2023), while RT-2 connects vision-language pretraining to robotic action prediction (Brohan et al., 2023). Open X-Embodiment aggregates demonstrations across robot platforms (Open X-Embodiment Collaboration et al., 2023), and Octo studies a generalist policy trained on this data and adapted to new settings (Octo Model Team et al., 2024). These works motivate scalable policy learning from diverse experience. Our policy experiment examines a different input to that experience: the physical asset library used to collect task demonstrations. It holds the policy family and demonstration count fixed across libraries, while differences in category match, trajectories, and asset preparation remain part of the training-source comparison.

3 Method

RFW-W combines the RFW-R rigid-asset pipeline, a common physical-asset interface, and task-conditioned scene composition (Figures 2 and 3).

3.1 Category-Guided Asset Generation

We use a human-curated taxonomy to control category coverage. qwen3.7-plus generates category-conditioned object descriptions; FLUX.2-klein (Black Forest Labs, 2026) and a supplementary Wan2.7 image path generate object images. RMBG-2.0 (BRIA AI, 2024) extracts foregrounds, and TRELLIS.2 (Xiang et al., 2025) reconstructs textured meshes. qwen3.7-plus (Yang et al., 2025a; Qwen Team, 2026) inspects five rendered views for completeness, recognizability, rigid-body suitability, and category agreement, then supports category reassignment and reinspection. The resulting RFW-R dataset contains 524,611 retained objects across 1,212 categories; the model roles and prompt specifications are in Appendix A.2.

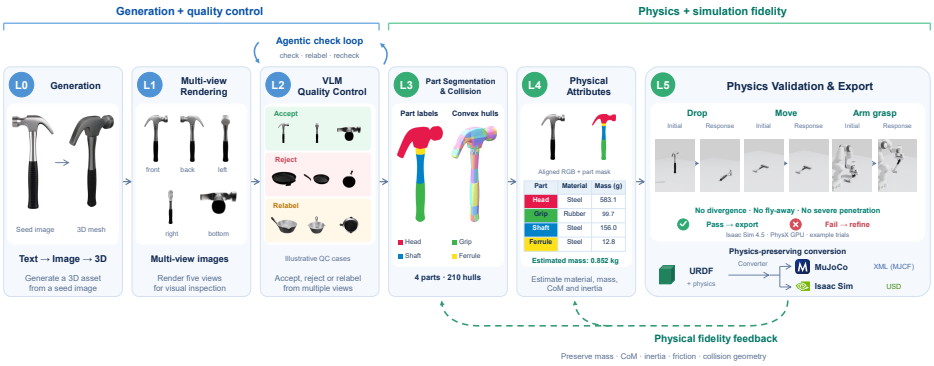


Figure 2: **RFW-R rigid-asset construction pipeline.** **Blue** stages (L0–L2) generate a 3D asset, render multiple views, and accept, reject, or relabel it through a quality-control loop. **Green** stages (L3–L5) assign semantic part labels and convex collision geometry, estimate physical attributes, and check simulated drop, motion, and arm-grasp behavior before export or refinement. In L3, the labeled colors distinguish semantic parts, while the multicolored facets show separate convex hulls. Solid arrows indicate forward processing; the blue loop indicates rechecking after relabeling, and the dashed green loop indicates feedback on physical fidelity. Validated assets support URDF, MJCF, and USD representations.

3.2 Physical and Interaction Representation

P3-SAM (Ma et al., 2025) identifies functional parts and CoACD (Wei et al., 2022) supplies convex collision geometry. qwen3.7-plus reads aligned RGB and part-mask views to estimate part materials

and friction, metric dimensions, and a whole-object mass prior. Parts carry material and interaction semantics; convex pieces define simulator contacts. Part masses follow $m_p = \rho_p V_p$ and are composed into object mass, center of mass, and inertia. When the composed-to-whole mass ratio falls outside $[0.5, 2]$, the whole-object estimate anchors the total while preserving relative part masses. Visual geometry, collision geometry, and inertial quantities share an object frame, and the resulting rigid body is exported to URDF and MJCF with part annotations retained for interaction queries. Appendix A.2 describes the annotation prompt, and Appendix A.4 gives the physical checks.

Interaction annotations connect semantic parts to executable grasps. We build on TRO-Grasp (Fei et al., 2025) and retain its hand-conditioned pose representation and multi-hand inference interface. Our extension conditions generation on the target semantic part through TRO's partial-point-cloud interface. Candidates are then filtered by part hit, real-mesh penetration, multi-finger engagement, closing, lifting, holding, and perturbation checks. We train each supported hand with the same asset-side annotation and filtering protocol, while preserving the visual mesh, collision representation, physical parameters, and task predicates. The candidate conversion, training flow, and execution checks appear in Appendix A.6.

3.3 Task-driven scene construction

RFW-W maps RFW-R metadata into a common description exposing the semantic, geometric, physical, and interaction fields required by each task. Retrieval checks these fields before binding an asset to a scene specification.

We first freeze a task core and then add the room and context without changing its internal arrangement (Figures 3 and 4). A language model compiles task semantics into a structured specification; retrieval and geometric solvers bind assets and place them. Functional checks and visual judgment are reported separately.

Figure 4 pairs task instructions with the corresponding scene visualizations across office, kitchen, and bedroom contexts. The examples illustrate qualitative correspondence between requested targets and scene content. Cross-engine RGB and surface-normal views are shown in Appendix A.5, Figure S7; object-level behavioral agreement is evaluated separately in Section 4.2.

Core assembly and room fitting. A task description is compiled into targets, supports, and required relations. Retrieval binds concrete assets through semantic relevance and physical compatibility, and a geometric solver assembles the relation-closed task core in a room-independent frame. The core membership and relative poses are then frozen. The room is fitted around this core, with walls, floors, doors, and windows generated by Infinigen (Raistrick et al., 2023; 2024); layout checks verify the planned boundaries and openings.

Appearance, packaging, and checks. Materials and peripheral furniture complete the scene subject to a protected-region constraint, and procedural materials are baked to PBR textures. The output is a structured description and self-contained USD package for Isaac Sim (NVIDIA, n.d.). Deterministic functional checks assess geometry, task constraints, and semantics; a qwen3.7-plus vision-language judge assesses realism and coherence. Failures enter separate functional or visual repair channels within bounded refinement through scene-repair levels R0–R5. These repair levels are distinct from the asset-construction stages L0–L5 in Figure 2. Refinement preserves the frozen core while functional and visual checks guide their respective repair channels.

4 Experiments

We evaluate three claims about RFW-R: its taxonomy supports broad rigid-asset coverage; its physical and interaction representations can be executed across simulators and hands; and its assets provide retrieval coverage, scene-population utility, and training data for downstream policies. Appendix A.1 summarizes these evaluation sets and protocols.

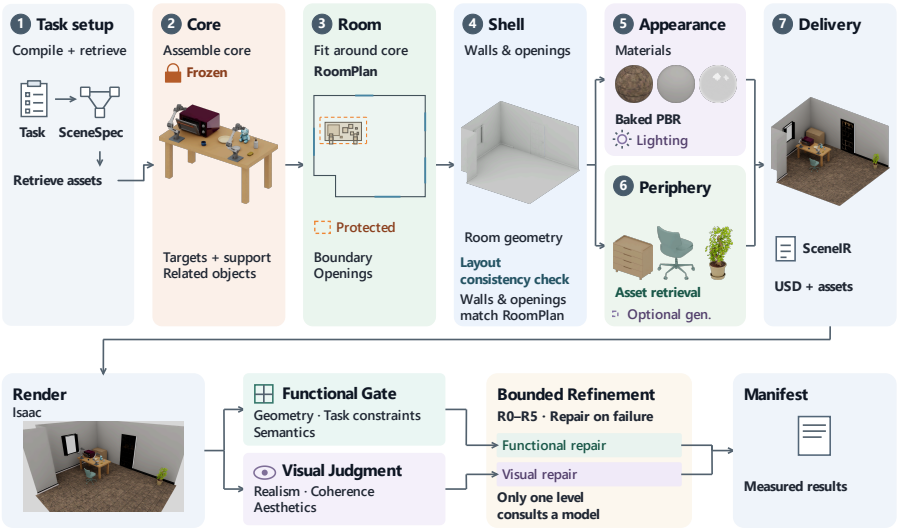


Figure 3: **Task-driven scene construction.** Task compilation and retrieval feed core assembly. The core is frozen before room fitting, shell construction, and packaging; functional and visual checks trigger bounded refinement.

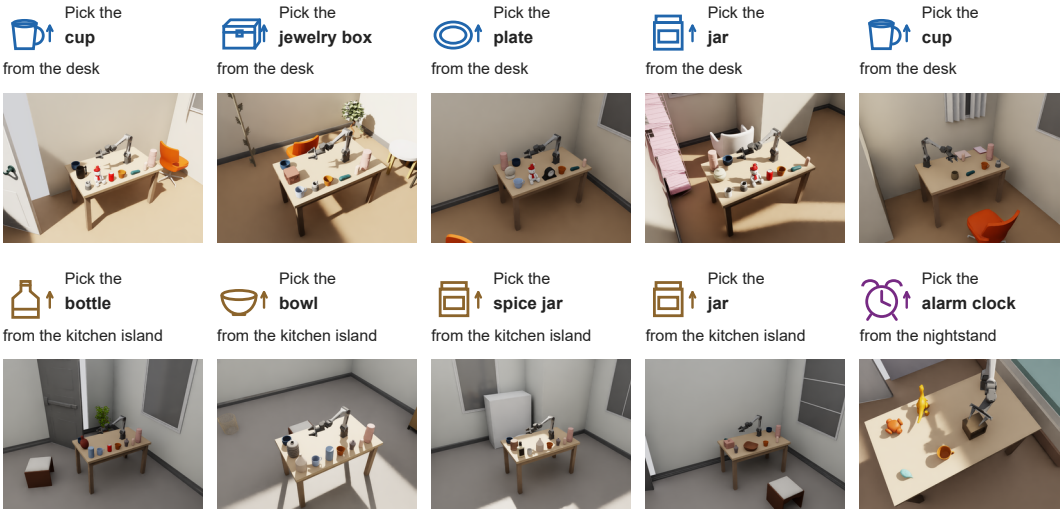


Figure 4: **Task-aligned scene examples.** Ten task instructions paired with their corresponding scene renderings. Each task is illustrated by an object pictogram and a description of the requested action, target object, and support surface.

4.1 Asset Annotation Quality

We evaluate semantic annotation agreement and physical-property consistency; metric definitions and object-level diagnostics are in Appendix A.3.

Under the shared vision-language protocol, RFW-R leads category (0.981) and functional (0.860) agreement, while its language score (0.864) is slightly below PhysX-3D (0.870) (Table 2).

Table 2: Part-level semantic annotation agreement under a shared vision-language protocol. Scores measure judge agreement; full diagnostics are in Appendix A.3.

Measure	RFW-R	EmbodiedGen v2	PhysX-3D	ManiTwin
Category ↑	**0.981**	0.93	0.96	0.65
Language ↑	0.864	0.67	**0.87**	0.50
Functional ↑	**0.860**	0.80	0.79	0.64

Physical representation. Across 2,424 assets, URDF/MJCF mass differs from the assigned mass by at most 1.5×10^{-6} relative error, and 98.5% pass the 10^{-3} tolerance between summed part masses and stored object mass. Appendix A.4 reports object-frame checks.

Effect of solid-fill modeling. On 31 retained assets, replacing the native representation with solid fill reduces the mean drop-response criterion-pass fraction by 9.25, 9.25, and 7.53 percentage points in MuJoCo, Genesis, and Isaac Sim, respectively. The native representation achieves higher pass fractions under the shared displacement, rotation, and penetration criteria across all three engines.

4.2 Cross-simulator drop-response consistency

We evaluate the same 2,257 assets in PyBullet (Coumans & Bai, 2016–2021), MuJoCo (Todorov et al., 2012), Genesis (Genesis Authors, 2024), and Isaac Sim (NVIDIA, n.d.) using estimated properties and multi-hull collisions. Each asset has three drop heights and five initial poses. The shared displacement, rotation, and penetration criteria and their aggregation are defined in Appendix A.5. Binary labels threshold each asset's criterion-pass fraction at 0.5, while Pearson r compares the continuous fractions.

Across 2,257 common assets, all four engines agree on the thresholded drop-response label for 80.77% of assets (Fleiss' $\kappa = 0.481$). Agreement includes both shared passes and shared failures under the same displacement, rotation, and penetration criteria. MuJoCo–Genesis shows the strongest agreement (94.20%, $\kappa = 0.741$) and correlation in per-asset criterion-pass fractions ($r = 0.919$) (Table 3). Evaluation details are provided in Appendix A.5.

Table 3: Cross-simulator drop-response consistency on 2,257 common assets. Agreement and κ use criterion-pass fractions binarized at 0.5; Pearson r uses continuous fractions. All-agree rate: 80.77% (Fleiss' $\kappa = 0.481$).

Engine pair	Agreement (%) ↑	Cohen's κ ↑	Pearson r ↑
PyBullet–MuJoCo	87.55	0.430	0.688
PyBullet–Genesis	86.71	0.398	0.670
PyBullet–Isaac Sim	87.90	0.331	0.650
MuJoCo–Genesis	94.20	0.741	0.919
MuJoCo–Isaac Sim	90.87	0.507	0.838
Genesis–Isaac Sim	89.68	0.450	0.830

4.3 Dexterous grasp execution

Six hand models execute close, lift, hold, and six-direction perturbation in Genesis. Lift success requires a rise above 0.075 m after a 0.15 m command; robust success additionally requires hold and perturbation across three mass/friction/pose settings. We report object-level any-success coverage from up to 64 candidates.

At the 64-candidate budget, mean coverage is 79.5% for lift and 73.7% for robust success. Ability and Inspire exceed 90% on both measures, while Shadow reaches 42.3% and 27.3%, demonstrating substantial hand-dependent variation under the shared evaluation pipeline (Table 4; Figure 5). Candidate-budget and contact diagnostics are in Appendix A.6.

4.4 Policy learning and real–simulation task views

We train separate $\pi_{0.5}$ (Physical Intelligence et al., 2025) policies for four libraries and five RoboTwin (Chen et al., 2025) tasks, using 100 demonstrations per library–task pair and checkpoint 19,999. Each policy is evaluated on 100 clean and 100 randomized expert-solvable scenes (4,000 rollouts); randomization changes visual context, table height, and language. Full settings are in Appendix A.7.

RFW-R reaches 71.2% clean and 41.8% randomized macro success, exceeding ManiTwin by 15.8 and 10.6 points. It leads on all clean tasks and four randomized tasks; PhysX-3D is higher on randomized Move Pillbottle Pad (9% versus 6%). The results show that the released assets support strong task performance under both nominal and varied scene conditions.

Table 4: Six-hand simulated grasp outcomes. Lift and robust rates use separate object sets; confidence intervals are in Appendix A.6.

Hand	N_L	Lift (%) ↑	N_R	Robust (%) ↑
Ability	2,372	92.5	2,367	90.4
Inspire	2,334	92.3	2,353	90.4
Allegro	2,372	91.3	2,372	81.6
Barrett	2,353	84.4	2,372	81.5
XHand	2,315	74.1	2,315	71.2
Shadow	2,353	42.3	2,372	27.3
Macro		79.5		73.7

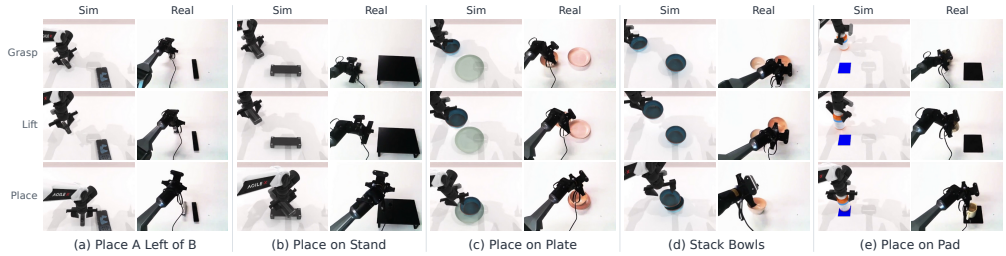


(a) Allegro Hand (b) Barrett Hand (c) Shadow Hand (d) Ability Hand (e) Inspire Hand (f) XHand

Figure 5: **Qualitative dexterous lifting across six hand morphologies.** Representative successful grasp-and-lift outcomes on released RFW-R assets under the Genesis lift criterion.

Table 5: Closed-loop RoboTwin success (%) after 100 demonstrations per library and task. Each cell has 100 expert-solvable scenes; macro averages weight the five tasks equally.

Mode	Training assets	A2B-L ↑	Stand ↑	Plate ↑	Bowls-2 ↑	Pillbottle-Pad ↑	Macro ↑
Clean	RFW-R	**47**	**63**	**96**	**98**	**52**	**71.2**
	EmbodiedGen v2	3	0	24	0	0	5.4
	PhysX-3D	22	51	93	42	6	42.8
	ManiTwin	33	50	81	80	33	55.4
Randomized	RFW-R	**21**	**33**	**76**	**73**	6	**41.8**
	EmbodiedGen v2	1	0	10	0	0	2.2
	PhysX-3D	8	16	48	24	**9**	21.0
	ManiTwin	8	29	56	59	4	31.2



(a) Place A Left of B (b) Place on Stand (c) Place on Plate (d) Stack Bowls (e) Place on Pad

Figure 6: **Real–simulation views of five RoboTwin tasks.** Simulated frames use the RoboTwin framework and rigid assets generated by the RFW-R pipeline; adjacent real-robot frames show grasp, lift, and place stages for the same task families.

The paired views in Figure 6 illustrate task roles, spatial relations, and grasp–lift–place stages across the same five RoboTwin task families in simulation and on the real robot.

4.5 Scene Construction and Fulfillment

We evaluate RFW-R, EmbodiedGen v2, and ManiTwin on 16 fixed room foundations: one living room, two kitchens, seven offices, two bedrooms, three bathrooms, and one dining room. The providers share object requests, target sizes, support relations, and placement rules. The 265 requested instances form 153 groups, with unresolved requests included in the fulfillment denominator. Camera selection uses shared target, azimuth, and distance candidates to maximize visibility

across providers. Each provider–foundation condition is rendered from five views, yielding 240 images across 48 conditions.

Table 6: Deterministic scene fulfillment on 16 fixed foundations. Fill counts are out of 265 instances and 153 groups; placement is the fraction successfully placed.

Library	Inst. fill ↑	Group fill ↑	Distinct ↑	Placement ↑
RFW-R	**265/265**	**153/153**	**80**	**100%**
EmbodiedGen v2	258/265	149/153	66	97.4%
ManiTwin	240/265	135/153	53	90.6%

RFW-R fills all 265 instances and 153 groups, compared with 258/265 and 149/153 for EmbodiedGen v2 and 240/265 and 135/153 for ManiTwin (Table 6). It also uses 80 distinct assets and places 100% of requested instances under the shared placement rules.

For visual assessment, a qwen3.7-plus vision-language judge scores each condition three times at temperature zero on realism, visual functionality, layout, completeness, prompt alignment, and style. Scores are averaged within each condition and then across the 16 foundations. Paired differences use 10,000 bootstrap resamples of foundations.

Table 7: Six-dimension visual assessment (0–10) across 48 provider–foundation conditions. Each condition is judged three times; entries average within conditions and then across 16 foundations.

Library	Realism	Visual func.	Layout	Complete	Prompt	Style	Mean
RFW-R	3.52	4.19	**4.88**	**3.02**	3.27	**4.40**	**3.878**
EmbodiedGen v2	**3.67**	**4.25**	4.71	2.92	**3.38**	4.15	3.844
ManiTwin	2.38	3.27	3.38	2.29	2.58	3.42	2.885

RFW-R has the highest mean visual score (3.878) and leads in layout, completeness, and style (Table 7). Its mean exceeds ManiTwin by 0.993 (95% bootstrap interval [0.611, 1.444]); the difference from EmbodiedGen v2 is 0.035 ($[-0.108, 0.191]$).

5 Conclusion

RFW-W provides a shared physical-asset interface and task-driven scene construction. The RFW-R rigid-object dataset contains 524,611 retained assets across 1,212 categories and fulfills all 265 requested instances across 16 fixed scene foundations. With 100 demonstrations per task and training source, RFW-R-trained policies reach 71.2% clean and 41.8% randomized macro success, exceeding the strongest baseline by 15.8 and 10.6 percentage points. Together, the inventory, scene, grasp, and policy results show the utility of RFW-R for embodied simulation and simulated policy learning.

AI USE STATEMENT

Generative AI tools, including OpenAI Codex, assisted with manuscript revision, information and reference search, reference formatting, figure-generation scripts, schematic task icons, and LaTeX layout. For the task-scene figure, the assistance composed supplied simulation renders with task descriptions from the accompanying metadata; the scene images were not generated or retouched during figure preparation. The research ideas, experimental code, experimental design, and interpretation of results were developed by the authors without generative-AI assistance.

Generative models are also components of the research pipeline itself: they generate object descriptions and synthetic object images, support asset inspection and semantic and physical annotation, and compile task descriptions into scene specifications. These roles are described in Section 3 and Appendix A.2. The authors are responsible for verifying AI-assisted text, code, figures, and references against the underlying sources and experimental records, and take full responsibility for the final manuscript and all reported results.

ETHICS STATEMENT

The quantitative evaluations concern generated rigid assets, simulated physical behavior, grasp execution, scene construction, and policy learning. The paper also includes recorded real-robot task views as qualitative illustrations (Figure 6 and Appendix A.7). These views do not establish real-robot success rates or deployment safety. Generated assets and model-derived annotations may reflect biases in the source models and category taxonomy. Estimated physical properties and simulation outcomes require validation for any intended physical deployment.

Third-party models, assets, robot descriptions, and simulator components remain subject to their respective terms and licenses. Our planned release will respect redistribution restrictions; for restricted components, we plan to provide source and version information and processing configurations in place of the restricted files.

REPRODUCIBILITY STATEMENT

Section 3 describes the construction and scene-composition pipelines. Appendix A.1 defines the inventory and evaluation cohorts; Appendices A.2–A.4 describe generation, annotation, and physical checks; Appendices A.5 and A.6 specify cross-engine and grasp-evaluation protocols; and Appendix A.7 records policy-training and evaluation settings, including the training seed, checkpoint, episode budgets, and aggregation rules. The policy results use one training seed per source–task pair. We plan to release dataset manifests, evaluation splits, scene specifications, metadata, training and evaluation code, model configurations, and scripts for the reported tables and figures, subject to redistribution rights. Recreating generated assets also depends on access to the model versions and services used in the construction pipeline.

REFERENCES

Black Forest Labs. FLUX.2 [klein]: Towards Interactive Visual Intelligence. https://bfl.ai/blog/flux2-klein-towards-interactive-visual-intelligence, January 2026.

BRIA AI. BRIA Background Removal v2.0 Model Card. https://huggingface.co/briaai/RMBG-2.0, 2024.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control, 2023. URL https://arxiv.org/abs/2307.15818.

Berk Calli, Aaron Walsman, Arjun Singh, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M. Dollar. Benchmarking in Manipulation Research: The YCB Object and Model Set and Benchmarking Protocols, 2015. URL https://arxiv.org/abs/1502.03143.

Ziang Cao, Zhaoxi Chen, Liang Pan, and Ziwei Liu. PhysX-3D: Physical-Grounded 3D Asset Generation, 2025a. URL https://arxiv.org/abs/2507.12465.

Ziang Cao, Fangzhou Hong, Zhaoxi Chen, Liang Pan, and Ziwei Liu. PhysX-Anything: Simulation-Ready Physical 3D Assets from Single Image, 2025b. URL https://arxiv.org/abs/2511.13648.

Ziang Cao, Yinghao Liu, Haitian Li, Runmao Yao, Fangzhou Hong, Zhaoxi Chen, Liang Pan, and Ziwei Liu. PhysX-Omni: Unified Simulation-Ready Physical 3D Generation for Rigid, Deformable, and Articulated Objects, 2026. URL https://arxiv.org/abs/2605.21572.

Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository, 2015. URL https://arxiv.org/abs/1512.03012.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, Weiliang Deng, Yubin Guo, Tian Nian, Xuanbing Xie, Qiangyu Chen, Kailun Su, Tianling Xu, Guodong Liu, Mengkang Hu, Huan-ang Gao, Kaixuan Wang, Zhixuan Liang, Yusen Qin, Xiaokang Yang, Ping Luo, and Yao Mu. RoboTwin 2.0: A Scalable Data Generator and Benchmark with Strong Domain Randomization for Robust Bimanual Robotic Manipulation, 2025. URL https://arxiv.org/abs/2506.18088.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion, 2023. URL https://arxiv.org/abs/2303.04137.

Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F. Yago Vicente, Thomas Dideriksen, Himanshu Arora, Matthieu Guillaumin, and Jitendra Malik. ABO: Dataset and Benchmarks for Real-World 3D Object Understanding, 2021. URL https://arxiv.org/abs/2110.06199.

Erwin Coumans and Yunfei Bai. PyBullet, a Python module for physics simulation for games, robotics and machine learning. https://github.com/bulletphysics/bullet3, 2016–2021.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, and Roozbeh Mottaghi. ProcTHOR: Large-Scale Embodied AI Using Procedural Generation, 2022. URL https://arxiv.org/abs/2206.06994.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-XL: A Universe of 10M+ 3D Objects, 2023a. URL https://arxiv.org/abs/2307.05663.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A Universe of Annotated 3D Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13142–13153, June 2023b. URL https://openaccess.thecvf.com/content/CVPR2023/html/Deitke_Objaverse_A_Universe_of_Annotated_3D_Objects_CVPR_2023_paper.html.

Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B. McHugh, and Vincent Vanhoucke. Google Scanned Objects: A High-Quality Dataset of 3D Scanned Household Items, 2022. URL https://arxiv.org/abs/2204.11918.

Clemens Eppner, Arsalan Mousavian, and Dieter Fox. ACRONYM: A Large-Scale Grasp Dataset Based on Simulation, 2020. URL https://arxiv.org/abs/2011.09584.

Xin Fei, Zhixuan Xu, Huaicong Fang, Tianrui Zhang, and Lin Shao. T(R,O) Grasp: Efficient Graph Diffusion of Robot-Object Spatial Transformation for Cross-Embodiment Dexterous Grasping, 2025. URL https://arxiv.org/abs/2510.12724.

Jiashi Feng, Xiu Li, Jing Lin, Jiahang Liu, Gaohong Liu, Weiqiang Lou, Su Ma, Guang Shi, Qinlong Wang, Jun Wang, Zhongcong Xu, Xuanyu Yi, Zihao Yu, Jianfeng Zhang, Yifan Zhu, Rui Chen, Jinxin Chi, Zixian Du, Li Han, Lixin Huang, Kaihua Jiang, Yuhan Li, Guan Luo, Shuguang Wang, Qianyi Wu, Fan Yang, Junyang Zhang, and Xuanmeng Zhang. Seed3D 1.0: From Images to High-Fidelity Simulation-Ready 3D Assets, 2025. URL https://arxiv.org/abs/2510.19944.

Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. LayoutGPT: Compositional Visual Planning and Generation with Large Language Models, 2023. URL https://arxiv.org/abs/2305.15393.

Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3D-FUTURE: 3D Furniture Shape with Texture, 2020. URL https://arxiv.org/abs/2009.09633.

Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, and Hao Zhang. 3D-FRONT: 3D Furnished Rooms With layOuts and semaNTics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10933–10942. IEEE, October 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Fu_3D-FRONT_3D_Furnished_Rooms_With_layOuts_and_semaNTics_ICCV_2021_paper.html.

Rao Fu, Zehao Wen, Zichen Liu, and Srinath Sridhar. AnyHome: Open-Vocabulary Generation of Structured and Textured 3D Homes, 2023. URL https://arxiv.org/abs/2312.06644.

Genesis Authors. Genesis: A Generative and Universal Physics Engine for Robotics and Beyond, December 2024. URL https://github.com/Genesis-Embodied-AI/genesis-world.

Haoran Geng, Helin Xu, Chengyang Zhao, Chao Xu, Li Yi, Siyuan Huang, and He Wang. GAPartNet: Cross-Category Domain-Generalizable Object Perception and Manipulation via Generalizable and Actionable Parts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7081–7091, June 2023.

URL https://openaccess.thecvf.com/content/CVPR2023/html/Geng_GAPartNet_Cross-Category_Domain-Generalizable_Object_Perception_and_Manipulation_via_Generalizable_and_CVPR_2023_paper.html.

Diandian Gu, Jing Lin, Gaohong Liu, Jiahang Liu, Su Ma, Guang Shi, Jun Wang, Qinlong Wang, Qianyi Wu, Zhongcong Xu, Xuanyu Yi, Zihao Yu, Jianfeng Zhang, Zhuolin Zheng, Yifan Zhu, Rui Chen, Hengkai Guo, Xiaoyang Guo, Mingcong Han, Xu Han, Xiu Li, Yixun Liang, Weiqiang Lou, Junzhe Lu, Guan Luo, Minghan Qin, Shuguang Wang, and Yuang Wang. Seed3D 2.0: Advancing High-Fidelity Simulation-Ready 3D Content Generation, 2026. URL https://arxiv.org/abs/2605.13862.

Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, Xiaodi Yuan, Pengwei Xie, Zhiao Huang, Rui Chen, and Hao Su. ManiSkill2: A Unified Benchmark for Generalizable Manipulation Skills, 2023. URL https://arxiv.org/abs/2302.04659.

Denys Iliash, Jiayi Liu, Egor Fokin, Qirui Wu, Ali Mahdavi-Amiri, Manolis Savva, and Angel X. Chang. Artiverse: A Diverse and Physically Grounded Dataset for Articulated Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8932–8942, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Iliash_Artiverse_A_Diverse_and_Physically_Grounded_Dataset_for_Articulated_Objects_CVPR_2026_paper.html.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, Aniruddha Kembhavi, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An Interactive 3D Environment for Visual AI, 2017. URL https://arxiv.org/abs/1712.05474.

Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, Andrey Kurenkov, C. Karen Liu, Hyowon Gweon, Jiajun Wu, Li Fei-Fei, and Silvio Savarese. iGibson 2.0: Object-Centric Simulation for Robot Learning of Everyday Household Tasks, 2021. URL https://arxiv.org/abs/2108.03272.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, Mona Anvari, Minjune Hwang, Manasi Sharma, Arman Aydin, Dhruva Bansal, Samuel Hunter, Kyu-Young Kim, Alan Lou, Caleb R Matthews, Ivan Villa-Renteria, Jerry Huayang Tang, Claire Tang, Fei Xia, Silvio Savarese, Hyowon Gweon, Karen Liu, Jiajun Wu, and Li Fei-Fei. BEHAVIOR-1K: A Benchmark for Embodied AI with 1,000 Everyday Activities and Realistic Simulation. In *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 80–93. PMLR, 14–18 Dec 2023. URL https://proceedings.mlr.press/v205/li23a.html.

Weiyu Li, Xuanyang Zhang, Zheng Sun, Di Qi, Hao Li, Wei Cheng, Weiwei Cai, Shihao Wu, Jiarui Liu, Zihao Wang, Xiao Chen, Feipeng Tian, Jianxiong Pan, Zeming Li, Gang Yu, Xiangyu Zhang, Daxin Jiang, and Ping Tan. Step1X-3D: Towards High-Fidelity and Controllable Generation of Textured 3D Assets, 2025. URL https://arxiv.org/abs/2505.07747.

Xian Li, Rong Wei, Lujie Yang, Haolin Huang, Junyuan Fang, Siliang Tang, Jun Xiao, Rui Tang, and Juncheng Li. UniPhysGen: Unified Physical Grounding for Simulation-Ready 3D Assets, 2026. URL https://arxiv.org/abs/2607.13586.

Chendi Lin, Heshan Liu, Qunshu Lin, Zachary Bright, Shitao Tang, Yihui He, Minghao Liu, Ling Zhu, and Cindy Le. Objaverse++: Curated 3D Object Dataset with Quality Annotations, 2025. URL https://arxiv.org/abs/2504.07334.

Chenguo Lin and Yadong Mu. InstructScene: Instruction-Driven 3D Indoor Scene Synthesis with Semantic Graph Prior, 2024. URL https://arxiv.org/abs/2402.04717.

Lu Ling, Chen-Hsuan Lin, Tsung-Yi Lin, Yifan Ding, Yu Zeng, Yichen Sheng, Yunhao Ge, Ming-Yu Liu, Aniket Bera, and Zhaoshuo Li. Scenethesis: A Language and Vision Agentic Framework for 3D Scene Generation, 2025. URL https://arxiv.org/abs/2505.02836.

Changfeng Ma, Yang Li, Xinhao Yan, Jiachen Xu, Yunhan Yang, Chunshi Wang, Zibo Zhao, Yanwen Guo, Zhuo Chen, and Chunchao Guo. P3-SAM: Native 3D Part Segmentation, 2025. URL https://arxiv.org/abs/2509.06784.

Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. MimicGen: A Data Generation System for Scalable Robot Learning using Human Demonstrations, 2023. URL https://arxiv.org/abs/2310.17596.

Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. PartNet: A Large-Scale Benchmark for Fine-Grained and Hierarchical Part-Level 3D Object Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 909–918, June 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Mo_PartNet_A_Large-Scale_Benchmark_for_Fine-Grained_and_Hierarchical_Part-Level_3D_CVPR_2019_paper.html.

Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. RoboCasa: Large-Scale Simulation of Everyday Tasks for Generalist Robots, 2024. URL https://arxiv.org/abs/2406.02523.

NVIDIA. Isaac Sim. https://developer.nvidia.com/isaac/sim/, n.d.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An Open-Source Generalist Robot Policy, 2024. URL https://arxiv.org/abs/2405.12213.

Open X-Embodiment Collaboration, Abby O'Neill, Abdul Rehman, et al. Open X-Embodiment: Robotic Learning Datasets and RT-X Models, 2023. URL https://arxiv.org/abs/2310.08864.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a Vision-Language-Action Model with Open-World Generalization, 2025. URL https://arxiv.org/abs/2504.16054.

Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, Vladimír Vondruš, Theophile Gervet, Vincent-Pierre Berges, John M. Turner, Oleksandr Maksymets, Zsolt Kira, Mrinal Kalakrishnan, Jitendra Malik, Devendra Singh Chaplot, Unnat Jain, Dhruv Batra, Akshara Rai, and Roozbeh Mottaghi. Habitat 3.0: A Co-Habitat for Humans, Avatars and Robots, 2023. URL https://arxiv.org/abs/2310.13724.

Qwen Team. Qwen3.7-Plus: Multimodal Agent Intelligence, May 2026. URL https://qwen.ai/blog?id=qwen3.7-plus.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, Alejandro Newell, Hei Law, Ankit Goyal, Kaiyu Yang, and Jia Deng. Infinite Photorealistic Worlds using Procedural Generation, 2023. URL https://arxiv.org/abs/2306.09310.

Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, Zeyu Ma, and Jia Deng. Infinigen Indoors: Photorealistic Indoor Scenes using Procedural Generation, 2024. URL https://arxiv.org/abs/2406.11824.

Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training Home Assistants to Rearrange their Habitat, 2021. URL https://arxiv.org/abs/2106.14405.

Stone Tao, Fanbo Xiang, Arth Shukla, Yuzhe Qin, Xander Hinrichsen, Xiaodi Yuan, Chen Bao, Xinsong Lin, Yulin Liu, Tse-kai Chan, Yuan Gao, Xuanlin Li, Tongzhou Mu, Nan Xiao, Arnav Gurha, Viswesh Nagaswamy Rajesh, Yong Woo Choi, Yen-Ru Chen, Zhiao Huang, Roberto Calandra, Rui Chen, Shan Luo, and Hao Su. ManiSkill3: GPU Parallelized Robotics Simulation and Rendering for Generalizable Embodied AI, 2024. URL https://arxiv.org/abs/2410.00425.

Team Hunyuan3D, Shuhui Yang, Mingxin Yang, Yifei Feng, Xin Huang, Sheng Zhang, Zebin He, Di Luo, Haolin Liu, Yunfei Zhao, Qingxiang Lin, Zeqiang Lai, Xianghui Yang, Huiwen Shi, Zibo Zhao, Bowen Zhang, Hongyu Yan, Lifu Wang, Sicong Liu, Jihong Zhang, Meng Chen, Liang Dong, Yiwen Jia, Yulin Cai, Jiaao Yu, Yixuan Tang, Dongyuan Guo, Junlin Yu, Hao Zhang, Zheng Ye, Peng He, Runzhou Wu, Shida Wei, Chao Zhang, Yonghao Tan, Yifu Sun, Lin Niu, Shirui Huang, Bojian Zheng, Shu Liu, Shilin Chen, Xiang Yuan, Xiaofeng Yang, Kai Liu, Jianchen Zhu, Peng Chen, Tian Liu, Di Wang, Yuhong Liu, Linus, Jie Jiang, Jingwei Huang, and Chunchao Guo. Hunyuan3D 2.1: From Images to High-Fidelity 3D Assets with Production-Ready PBR Material, 2025. URL https://arxiv.org/abs/2506.15442.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109.

Boyuan Wang, Yue Zhang, Xutao Xue, Xueyu Song, and Yu Sun. TableVerse: A Large-scale Tabletop Dataset with Real-world Grounded Layouts for Generalizable Manipulation, 2026a. URL https://arxiv.org/abs/2607.21017.

Kaixuan Wang, Tianxing Chen, Jiawei Liu, Honghao Su, Shaolong Zhu, Minxuan Wang, Zixuan Li, Yue Chen, Huan-ang Gao, Yusen Qin, Jiawei Wang, Qixuan Zhang, Lan Xu, Jingyi Yu, Yao Mu, and Ping Luo. ManiTwin: Scaling Data-Generation-Ready Digital Object Dataset to 100K, 2026b. URL https://arxiv.org/abs/2603.16866.

Lirui Wang, Yiyang Ling, Zhecheng Yuan, Mohit Shridhar, Chen Bao, Yuzhe Qin, Bailin Wang, Huazhe Xu, and Xiaolong Wang. GenSim: Generating Robotic Simulation Tasks via Large Language Models, 2023a. URL https://arxiv.org/abs/2310.01361.

Ruicheng Wang, Jialiang Zhang, Jiayi Chen, Yinzhen Xu, Puhao Li, Tengyu Liu, and He Wang. DexGraspNet: A Large-Scale Robotic Dexterous Grasp Dataset for General Objects Based on Simulation, 2022. URL https://arxiv.org/abs/2210.02697.

Xinjie Wang, Liu Liu, Yu Cao, Ruiqi Wu, Wenkang Qin, Dehui Wang, Wei Sui, and Zhizhong Su. EmbodiedGen: Towards a Generative 3D World Engine for Embodied Intelligence, 2025. URL https://arxiv.org/abs/2506.10600.

Xinjie Wang, Liu Liu, Taojun Ding, Andrew Choi, Chaodong Huang, Mengao Zhao, Ziang Li, Jackson Jiang, Chunlei Yu, Shengxiang Liu, Wei Xu, and Zhizhong Su. EmbodiedGen V2: An Agentic, Simulation-Ready 3D World Engine for Embodied AI, 2026c. URL https://arxiv.org/abs/2607.07459.

Yufei Wang, Zhou Xian, Feng Chen, Tsun-Hsuan Wang, Yian Wang, Katerina Fragkiadaki, Zackory Erickson, David Held, and Chuang Gan. RoboGen: Towards Unleashing Infinite Data for Automated Robot Learning via Generative Simulation, 2023b. URL https://arxiv.org/abs/2311.01455.

Xinyue Wei, Minghua Liu, Zhan Ling, and Hao Su. Approximate Convex Decomposition for 3D Meshes with Collision-Aware Concavity and Tree Search, 2022. URL https://arxiv.org/abs/2205.02961.

Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, Dahua Lin, and Ziwei Liu. OmniObject3D: Large-Vocabulary 3D Object Dataset for Realistic Perception, Reconstruction and Generation, 2023. URL https://arxiv.org/abs/2301.07525.

Hongchi Xia, Xuan Li, Zhaoshuo Li, Qianli Ma, Jiashu Xu, Ming-Yu Liu, Yin Cui, Tsung-Yi Lin, Wei-Chiu Ma, Shenlong Wang, Shuran Song, and Fangyin Wei. SAGE: Scalable Agentic 3D Scene Generation for Embodied AI, 2026. URL https://arxiv.org/abs/2602.10116.

Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao Su. SAPIEN: A SimulAted Part-Based Interactive ENvironment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11097–11107. IEEE, June 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Xiang_SAPIEN_A_SimulAted_Part-Based_Interactive_ENvironment_CVPR_2020_paper.html.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3D Latents for Scalable and Versatile 3D Generation, 2024. URL https://arxiv.org/abs/2412.01506.

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, 2025. URL https://arxiv.org/abs/2512.14692.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025a. URL https://arxiv.org/abs/2505.09388.

Yandan Yang, Baoxiong Jia, Peiyuan Zhi, and Siyuan Huang. PhyScene: Physically Interactable 3D Scene Synthesis for Embodied AI, 2024. URL https://arxiv.org/abs/2404.09465.

Yandan Yang, Baoxiong Jia, Shujie Zhang, and Siyuan Huang. SceneWeaver: All-in-One 3D Scene Synthesis with an Extensible and Self-Reflective Agent, 2025b. URL https://arxiv.org/abs/2509.20414.

Yi Yang, Hao Pan, Yijing Cui, Alla Sheffer, and Changjian Li. Sketch2Arti: Sketch-based Articulation Modeling of CAD Objects. *ACM Transactions on Graphics*, 45(4), 2026a. doi: 10.1145/3811375. URL https://arlo-yang.github.io/Sketch2Arti/.

Yue Yang, Fan-Yun Sun, Luca Weihs, Eli VanderBilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, Chris Callison-Burch, Mark Yatskar, Aniruddha Kembhavi, and Christopher Clark. Holodeck: Language Guided Generation of 3D Embodied AI Environments, 2023. URL https://arxiv.org/abs/2312.09067.

Yunhan Yang, Chunshi Wang, Junliang Ye, Yang Li, Zanxin Chen, Zehuan Huang, Yao Mu, Zhuo Chen, Chunchao Guo, and Xihui Liu. PhysForge: Generating Physics-Grounded 3D Assets for Interactive Virtual World, 2026b. URL https://arxiv.org/abs/2605.05163.

Hui Zhang, Sammy Christen, Zicong Fan, Otmar Hilliges, and Jie Song. GraspXL: Generating Grasping Motions for Diverse Objects at Scale, 2024a. URL https://arxiv.org/abs/2403.19649.

Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. CLAY: A Controllable Large-scale Generative Model for Creating High-quality 3D Assets, 2024b. URL https://arxiv.org/abs/2406.13897.

Yang Zhao, Shizhao Sun, Meisheng Zhang, Yingdong Shi, Xubo Yang, and Jiang Bian. SceneReVis: A Self-Reflective Vision-Grounded Framework for 3D Indoor Scene Synthesis via Multi-turn RL, 2026. URL https://arxiv.org/abs/2602.09432.

APPENDIX CONTENTS

Detailed methods, evaluation protocols, and supplementary results.

A.1 Datasets and Evaluation Sets . . . 19
Dataset coverage . . . 19
Category-balanced evaluation . . . 20
Release scope and within-category diversity . . . 20
A.2 Asset Construction Details . . . 22
Generation and category reassignment . . . 22
Model and prompt specifications . . . 22
Physicalization pipeline . . . 22
Geometry cleanup . . . 22
Recorded processing cost . . . 23
A.3 Geometry and Semantic Evaluation . . . 23
Geometry metrics . . . 23
Semantic annotation metrics . . . 23
Paired geometry and collision fidelity . . . 24
A.4 Physical Representation and Composition . . . 24
Part-level physical properties . . . 24
Physical consistency . . . 25
Effect of solid-fill modeling . . . 25
Appearance, parts, and collision geometry . . . 26
A.5 Cross-Simulator Evaluation . . . 26
Drop-response protocol . . . 26
Agreement metrics . . . 26
A.6 Dexterous Grasping: Protocol and Diagnostics . . . 27
Candidate generation and execution . . . 27
Candidate budgets . . . 28
Robustness and closure displacement . . . 29
A.7 Policy Data, Training and Evaluation . . . 29
Data budget and asset selection . . . 29
Training and evaluation settings . . . 29
Aggregation and task-level results . . . 30
Task-semantic correspondence between simulation and real setups . . . 30

A.1 Datasets and Evaluation Sets

A.1.1 Inventory and evaluation scope

The RFW-R rigid-object dataset contains 524,611 assets across 1,212 leaf categories. Table S1 compares it with EmbodiedGen v2 (Wang et al., 2026c), PhysX-3D (Cao et al., 2025a), and ManiTwin (Wang et al., 2026b). We report entropy-equivalent coverage $N_{\text{eff}} = \exp(-\sum_c p_c \log p_c)$ within each library's taxonomy, where $p_c = n_c / \sum_j n_j$. The balanced evaluation samples two eligible assets per category (2,424 objects), independently of simulation and manipulation outcomes.

Table S1: Evaluated release inventories. N_{eff} is entropy-equivalent class coverage within each library taxonomy.

Dataset	Assets	Classes	N_{eff}	Med./class
RFW-R	**524,611**	**1,212**	**905.1**	**356**
EmbodiedGen v2	4,168	200	139.2	11
PhysX-3D	32,047	450	7.8	1
ManiTwin	4,574	485	285.6	6

A.1.2 Dataset coverage

RFW-R contains 524,611 retained rigid assets in 1,212 categories. The comparison uses EmbodiedGen v2, PhysX-3D, and ManiTwin as the asset sources specified in the main paper.

Figure S1 summarizes the breadth of the rigid catalog with examples from eight functional groups and descriptive distributions across domains, classes, and estimated physical properties.



Figure S1: **RFW-R dataset coverage.** Top: 48 representative assets from eight functional groups, individually scaled for visibility. Bottom: rounded object counts across 30 domains, a frequency-weighted word cloud of 1,212 categories, and distributions of estimated mass, metric height, and part materials. The final RFW-R release contains 524,611 retained assets across 1,212 categories.

A.1.3 Category-balanced evaluation

We select two objects from each category with available geometry and physical descriptions, giving 2,424 objects. Selection does not depend on simulation or manipulation outcomes. The categories span 30 domains and comprise 121 head, 364 middle, and 727 tail categories, contributing 242, 728, and 1,454 selected objects.

Table S2: Evaluation settings.

Evaluation	Size	Unit	Setting
Rigid inventory	524,611	assets	1,212 categories
Category-balanced sample	2,424	assets	Two per category
Four-engine agreement	2,257	assets	Common evaluable subset
Policy training	2,000	demonstrations	100 per source and task
Policy evaluation	4,000	episodes	100 per policy and condition
Scene construction	16	foundations	265 instances, three providers

A.1.4 Release scope and within-category diversity

The retained inventory contains 524,611 assets; simulation export completes for 524,562 of them (99.99%). All 2,424 category-balanced evaluation assets have complete simulation representations.

On the 2,424-object subset, an image-feature screen identifies 24 candidate near-duplicate clusters involving 55 objects (2.27%) at cosine similarity 0.9782. The threshold is calibrated with same-category and cross-category proxy pairs to select candidates for visual inspection. Across the 1,212

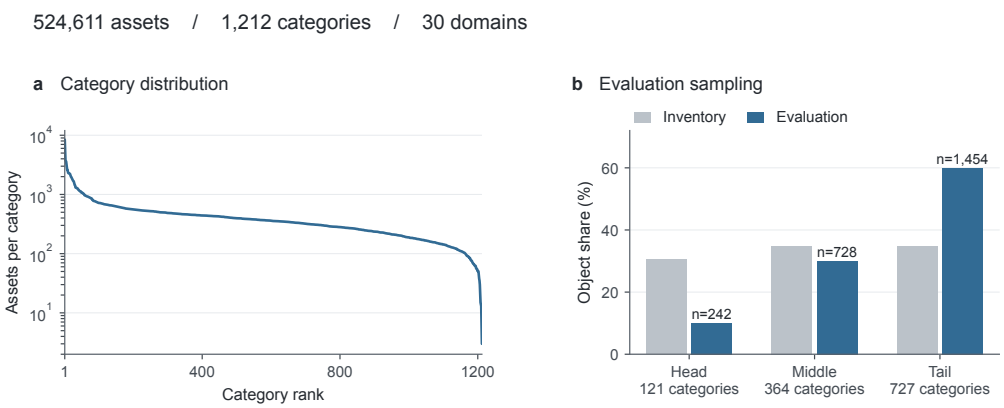


Figure S2: **Rigid inventory and category-balanced sampling.** The category population curve uses a logarithmic vertical axis. Bars compare inventory and evaluation shares; n labels give evaluation object counts. Sampling selects two objects per category.

within-category pairs, image-feature distance has median 0.1108 and 5th–95th percentiles 0.0525–0.1764.

A.2 Asset Construction Details

A.2.1 Generation and category reassignment

Image generation, background removal, mesh reconstruction, five-view inspection, and category reassignment produce the retained inventory. The visual rubric checks single-object content, rigid-body suitability, completeness, floating fragments, recognizability, and category agreement. Objects failing only category agreement are reassigned through lexical and image-conditioned matching, followed by reinspection for breakage and multi-object assemblies. Suitable unmatched objects form new size-bounded classes.

Table S3: Candidate generation and retention after visual filtering and category reassignment.

Stage or disposition	Assets
Generated candidates	599,439
Retained with original label	357,117
Retained after reclassification	147,573
Retained in 12 induced categories	19,921
Total retained	524,611
Not retained in this inventory	74,828

A.2.2 Model and prompt specifications

Category-conditioned generation. qwen3.7-plus receives the domain and the English and Chinese category names and produces batches of English object descriptions in JSON. The prompt requests one sentence of 15–30 words per description, varying object sub-form, material and finish, style and era, color and pattern, condition and wear, and notable geometry. Twelve rotating diversity cues specify combinations of materials, styles, conditions, and colors. Generation uses temperature 0.9 and top-p 0.95. Descriptions are deduplicated by token-set Jaccard similarity at 0.85 and screened for domain-specific material plausibility. FLUX.2-klein renders images from the selected descriptions with six inference steps and guidance scale 1.0; the supplementary Wan2.7 path samples descriptions across each category.

Visual quality and category prompts. qwen3.7-plus (Yang et al., 2025a; Qwen Team, 2026) receives front, back, left, right, and bottom renders. Its ordered rubric asks whether the asset is a single object, a rigid body, complete, free of floating fragments, recognizable, and consistent with the requested domain and category. The response records these decisions and an object guess used for category reassignment. Ambiguous assignments are resolved from the rendered views and a candidate-category list, with an abstention choice. Follow-up prompts score visible breakage and the number of separate objects in a render.

Physical annotation prompt. qwen3.7-plus receives aligned textured RGB and color-coded part-mask views. The prompt uses the mask colors to locate parts and the RGB views to identify material category and subtype. For each part, it requests static and dynamic friction against rubber; for the complete rigid object, it requests metric dimensions and a mass estimate. The prompt asks for one answer per mask color and treats small or occluded parts in the context of the whole object. Material densities and part volumes then determine the composed mass, center of mass, and inertia in Appendix A.4.

A.2.3 Physicalization pipeline

Figure 2 shows the detailed construction stages. Segmentation and collision decomposition establish the physical representation; estimated material and inertial properties and simulation checks connect that representation to downstream execution.

A.2.4 Geometry cleanup

Conservative cleanup targets coincident vertices, fragmented components, and inconsistent part annotations while preserving openings and graspable structures. Category reassignment and geometric

cleanup serve different purposes: the former recovers category coverage, while the latter improves the representation used for annotation and export. A paired audit of the construction stages finds that visual surfaces remain essentially unchanged after coordinate alignment; Appendix A.3.3 reports the measurements. Physical normalization aligns the physical properties and geometry in a common object frame while preserving the visual surface.

Three independent examples illustrate category reassignment, a remaining open boundary after cleanup, and a drop-response criterion failure (Figure S3).

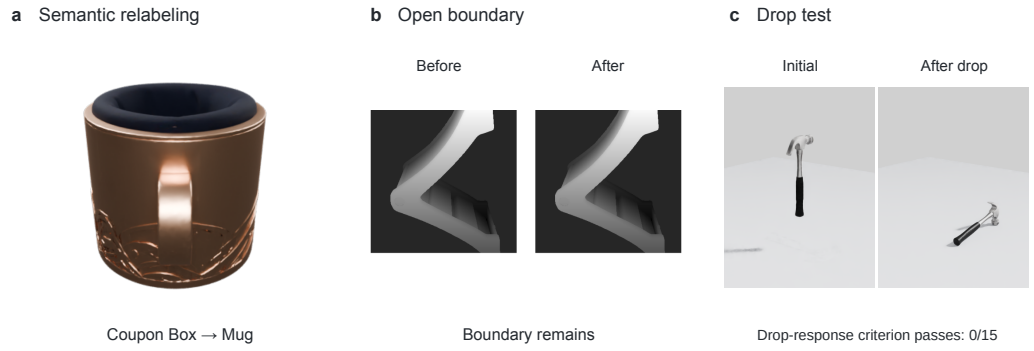


Figure S3: **Observed diagnostic cases.** A source category is reassigned from Coupon Box to Mug; a boundary remains open after processing; and a claw hammer fails the specified drop-response criterion in all 15 archived drop trials. The panels show three independent diagnostic examples.

A.2.5 Recorded processing cost

Table S4 reports CPU time for physical normalization, repair, and collision reconstruction. Each stage is measured independently on the asset set listed in the table.

Table S4: Recorded asset-processing cost. CPU hours denote accumulated compute time. Asset counts correspond to the available timing records for each stage.

Stage	Recorded assets	CPU hours	Mean seconds/asset
Physical normalization	524,611	10,723.2	73.6
Repair	2,424	35.4	52.6
Collision reconstruction	1,291	358.8	1,000.5

A.3 Geometry and Semantic Evaluation

A.3.1 Geometry metrics

We characterize connected components, boundary-edge ratio, visual volume relative to convex-hull volume, and visual-to-collision surface distances. Deep intrusion measures collision samples lying more than 1 cm inside the visual volume and is evaluated on watertight meshes. The original diagnostic settings use a component threshold of 5,000, boundary ratio 0.30, solidity 0.20, visual-to-collision P95 distance 2 cm, and intrusion fraction 0.02.

The baseline conversion for PhysX-3D and ManiTwin uses visual geometry for collision, so its visual-to-collision distance is a self-comparison. Geometry scores therefore depend on the collision representation as well as the source asset. Table S5 reports the diagnostic results for these source-specific collision representations.

A.3.2 Semantic annotation metrics

The four libraries share a qwen3.7-plus vision-language scoring and aggregation protocol for category, language, functional, and physics-related agreement. CLIP (Radford et al., 2021) image–text similarity provides an additional alignment measure. The physics-related judge score measures

Table S5: Original geometry diagnostic pass rates (%). Collision representations differ across sources as described in the text.

Gate (pass rate ↑)	RFW-R	EmbodiedGen v2	PhysX-3D	ManiTwin
Geometry-only strict	79.3	71.1	53.7	**89.5**
Connectivity	98.6	98.2	100	100
Open-boundary edges	98.9	96.1	99.9	100
Visual→collision dist.	98.8	96.4	100	100
Deep penetration	99.5	100	100	100
Hull solidity	81.6	75.0	53.8	**89.5**

agreement with the estimated physical annotations. N/A indicates that the released comparison sources do not provide a comparable physics-judge field under this protocol.

Table S6: Part-level annotation quality under the shared judge. All four libraries use the same scoring and aggregation protocol. *EmbodiedGen v2 physics is computed on placeholder values; N/A denotes an unavailable comparable field.

Dimension	RFW-R	EmbodiedGen v2	PhysX-3D	ManiTwin
Category ↑	**0.981**	0.93	0.96	0.65
Language ↑	0.864	0.67	**0.87**	0.50
Functional ↑	**0.860**	0.80	0.79	0.64
Physics (judge) ↑	**0.895**	0.85*	N/A	N/A
CLIP alignment ↑	0.244	0.232	0.238	**0.261**

A.3.3 Paired geometry and collision fidelity

A paired audit tracks 2,424 assets through geometry cleanup and physical normalization. After coordinate alignment, the median per-object bidirectional surface-distance P95 is 3.87×10^{-7} m before and after physical normalization. Watertightness is approximately 10.1% before and after normalization, consistent with preservation of the visual surface.

A separate stratified sample of 200 assets compares the released collision geometry, standard CoACD, and a single convex hull. We compute volume IoU on the 25 objects with valid closed visual-volume proxies. On this measurable subset, published collisions and standard CoACD both attain higher median IoU than a single hull (Table S7).

Table S7: Collision fidelity on the same 25 measurable objects from a 200-asset sample. IoU compares each collision volume with the closed visual-volume proxy.

Collision representation	Measurable objects	Median volume IoU ↑
Released representation	25	**0.924**
Standard CoACD	25	<u>0.913</u>
Single convex hull	25	0.682

Annotation continuity. Affordance annotations and the inputs to semantic evaluation are unchanged before and after physical normalization for all 2,424 audited objects. This checks continuity during physical normalization.

A.4 Physical Representation and Composition

A.4.1 Part-level physical properties

Each part associates metric geometry with material, density, friction, and semantic attributes. Given part masses $m_p = \rho_p V_p$, the object mass and center of mass are

$$M = \sum_p m_p, \qquad c = \frac{1}{M} \sum_p m_p c_p. \tag{1}$$

Part inertia tensors are combined in the object frame using the parallel-axis theorem. A whole-object mass estimate $\widehat{M}$ anchors all part masses by the common factor $\widehat{M}/M$ when $M/\widehat{M} \notin [0.5, 2]$. Uniform scaling preserves the composed center of mass and relative part masses.

Figure S4 compares assigned part masses with a uniform-density allocation at the same total object mass. This isolates how the part annotations redistribute mass within each asset.

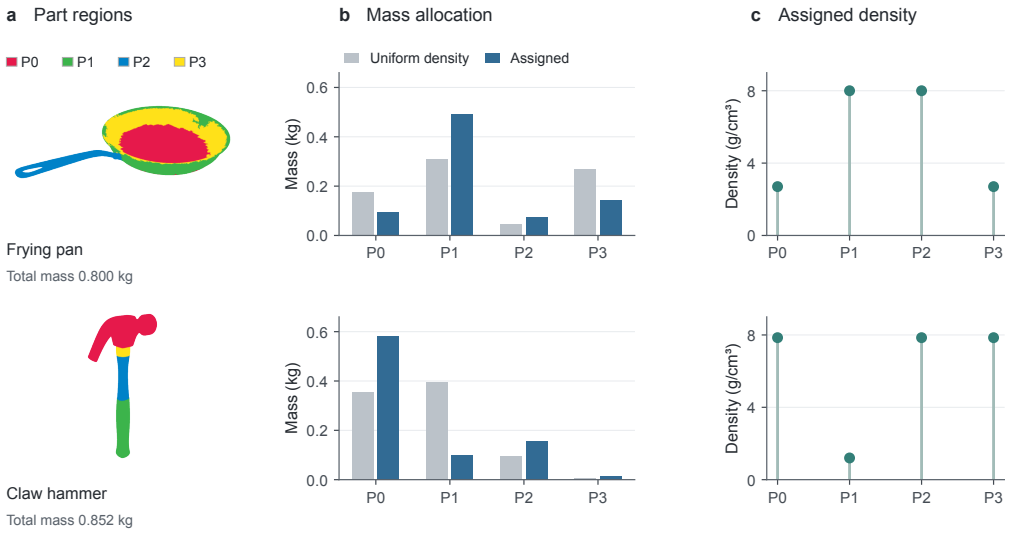


Figure S4: **Part-wise physical assignment.** For each asset, the uniform-density baseline allocates the fixed total mass in proportion to stored part volume. Assigned masses use the exported annotations; both allocations sum to 0.800 kg for the pan and 0.852 kg for the hammer. The comparison shows how part annotations redistribute mass within the common object representation.

A.4.2 Physical consistency

We evaluate the consistency of physical representations on 2,424 assets. Object-level quantities are transformed with the geometry into a common frame; all audited center-of-mass and inertia tensors agree after alignment to the common object frame. Object mass recovered from URDF and MJCF has maximum relative error 1.5×10^{-6}. Part-mass sums agree with stored object mass within relative tolerance 10^{-3} for 98.5% of assets. Preserving these physical quantities across simulator exports supports the reuse of RFW-R assets in scene construction and manipulation policy learning.

Table S8: Physical-representation checks on 2,424 assets.

Check	Result
Simulation-description validity	2,424/2,424
Object-frame COM and inertia consistency	2,424/2,424
URDF/MJCF mass relative error (maximum) ↓	1.5×10^{-6}
Part-sum mass agreement within 10^{-3} ↑	98.5%

A.4.3 Effect of solid-fill modeling

We compare the native representation with solid fill on 31 retained assets. One asset removed from the asset library because of a degenerate mass assignment is excluded from both conditions in every engine. The comparison uses MuJoCo, Genesis, and Isaac Sim, with three drop heights and five poses per asset, 2.5-second rollouts, and a 0.01-second time step. The two conditions yield 2,790 trials with no execution failures. Trial pass/fail labels follow the drop-response criteria in Appendix A.5. Each asset contributes a criterion-pass fraction over 15 trials, and paired differences are aggregated across assets.

Table S9 isolates the solid-fill comparison. The native representation achieves higher criterion-pass fractions than solid fill in all three engines under the shared displacement, rotation, and penetration criteria.

Table S9: Effect of solid fill on 31 retained assets. Changes are solid fill minus native in criterion-pass fraction, expressed in percentage points. Approximate per-comparison 95% intervals are computed as the paired mean ± 1.96 standard errors across assets.

Engine	Mean change (pp)	Approx. 95% interval
MuJoCo	-9.25	$[-16.01, -2.48]$
Genesis	-9.25	$[-15.88, -2.62]$
Isaac Sim	-7.53	$[-13.52, -1.54]$

A.4.4 Appearance, parts, and collision geometry

A single rigid body can contain multiple semantic parts. Figure S5 aligns the appearance, part segmentation, and collision representation of two assets. The frying pan and claw hammer each contain four annotated parts, with 92 and 210 convex hulls, respectively, in these illustrative part-based decompositions. The quantitative experiments use whole-object collision decomposition.



Figure S5: **Aligned asset representations.** Appearance, original part labels, convex collision pieces, and assigned masses are shown for a frying pan and claw hammer. Part colors identify P0–P3; collision colors distinguish convex pieces. The masses are exported estimates.

A.5 Cross-Simulator Evaluation

A.5.1 Drop-response protocol

We evaluate the same 2,257 assets in all four engines. Each object is evaluated from heights 0.05, 0.15, and 0.30 m with five initial orientations. The rollout lasts 2.5 seconds with a 0.01-second step. A valid trial passes the drop-response criterion when horizontal displacement is at most $\max(0.05\,\mathrm{m}, 0.1d)$, rotation change at most $8°$, and support penetration no deeper than 0.01 m, where d is the object size. All engines use the estimated physical properties and multi-hull collision geometry.

A.5.2 Agreement metrics

For asset i and engine e, let s_{ie} be the fraction of valid trials that pass all three drop-response criteria and $b_{ie} = \mathbb{1}[s_{ie} \geq 0.5]$. Pairwise agreement compares b_{ie}; Cohen's κ corrects for marginal label frequencies. Fleiss' κ summarizes four-engine agreement, and Pearson correlation uses the continuous values s_{ie}.

All pairs use the same 2,257 assets. All-engine agreement is 80.77% with Fleiss' $\kappa = 0.481$. Positive asset labels ($b_{ie} = 1$) account for 7.75–13.03% across engines. Agreement counts both matching positive and matching negative labels.

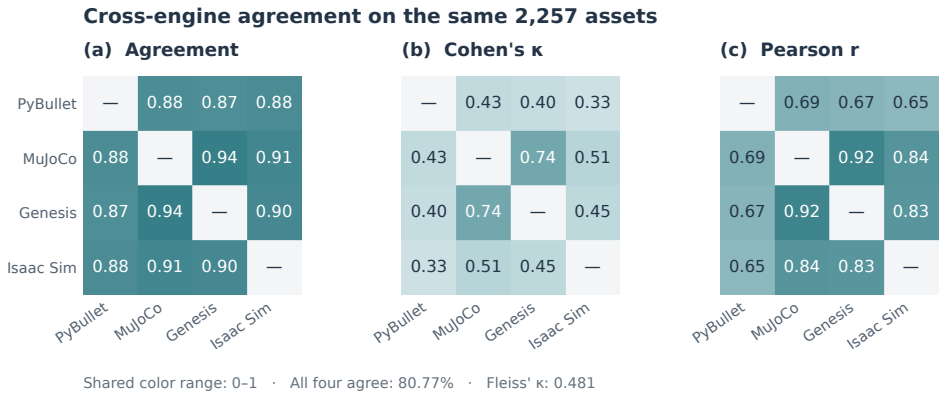


Figure S6: **Drop-response consistency across four physics engines.** Pairwise agreement, Cohen's κ, and criterion-pass-fraction correlation use the same 2,257 assets. Values are calculated from the evaluated trials.

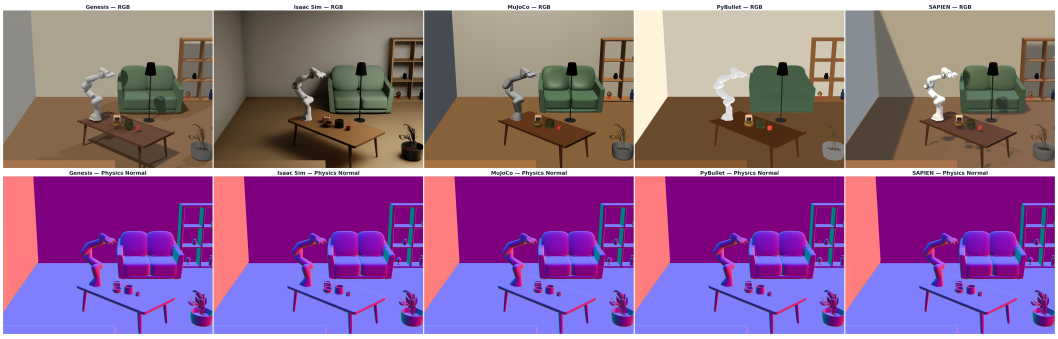


Figure S7: **Cross-simulator scene rendering.** The same composed scene is rendered in five physics engines through the shared asset interface. Top: RGB views. Bottom: surface-normal views. These views illustrate scene structure and engine-dependent appearance through the shared interface. Task–scene correspondence is shown separately in Figure 4; behavioral agreement is evaluated by the drop-response protocol above.

A.6 Dexterous Grasping: Protocol and Diagnostics

A.6.1 Candidate Generation and Execution

We build on T(R,O) Grasp (TRO-Grasp) (Fei et al., 2025), which generates hand-conditioned object–robot transformations. For each target asset, we retain the generated pose representation and condition the candidate pool on the selected semantic part through the partial point-cloud interface. Candidates are converted to the common asset frame, checked against the semantic part and real mesh, and then passed to the same execution protocol for every trained hand. Up to 64 candidates per object–hand pair are evaluated. The hand pose comprises root translation/orientation and joint configuration.

Genesis executes closing, a 0.15 m commanded lift, holding and six-direction perturbation under open-loop position control. Lift-only success requires object rise greater than 0.075 m. Object-level any-success counts an object as successful when at least one of its evaluated candidates succeeds. The robust records combine three mass/friction/pose settings. They use different cohorts from the lift-only records, as shown explicitly below.

A saved execution example in Figure S8 shows the resulting close–lift–hold sequence.

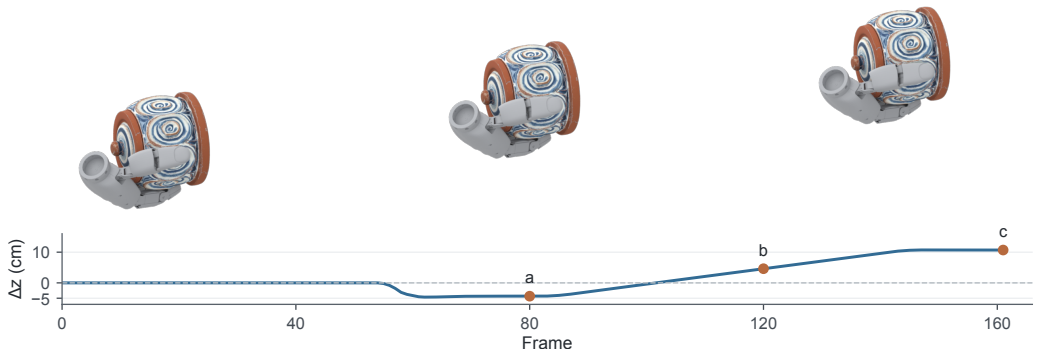


Figure S8: **Example grasp execution.** XHand closes, lifts, and holds a condiment jar in three saved poses. Markers a–c locate frames 80, 120, and 161 on the vertical-displacement curve, measured relative to frame 0. The sequence illustrates the close, lift, and hold stages of object manipulation.

A.6.2 CANDIDATE BUDGETS

The candidate budget controls how much of the generated grasp distribution is exposed to execution. Figure S9 shows the seven recorded budgets: coverage rises as the budget grows, with the largest gains for several hands occurring between 16 and 64 candidates. Table S10 gives the full-budget rates and the separate robust cohorts.

Table S10: Dexterous grasp outcomes in Genesis at the full candidate budget. Lift and robust rates use separate object cohorts (N_L, N_R). Brackets give Wilson 95% intervals; macro averages include all six hands.

Hand	N_L	Lift [95% CI] ↑	N_R	Robust ↑
Ability	2,372	92.5% [91.3, 93.4]	2,367	90.4%
Inspire	2,334	92.3% [91.1, 93.3]	2,353	90.4%
Allegro	2,372	91.3% [90.1, 92.4]	2,372	81.6%
Barrett	2,353	84.4% [82.8, 85.8]	2,372	81.5%
XHand	2,315	74.1% [72.3, 75.8]	2,315	71.2%
Shadow	2,353	42.3% [40.3, 44.3]	2,372	27.3%
Macro		79.5%		73.7%

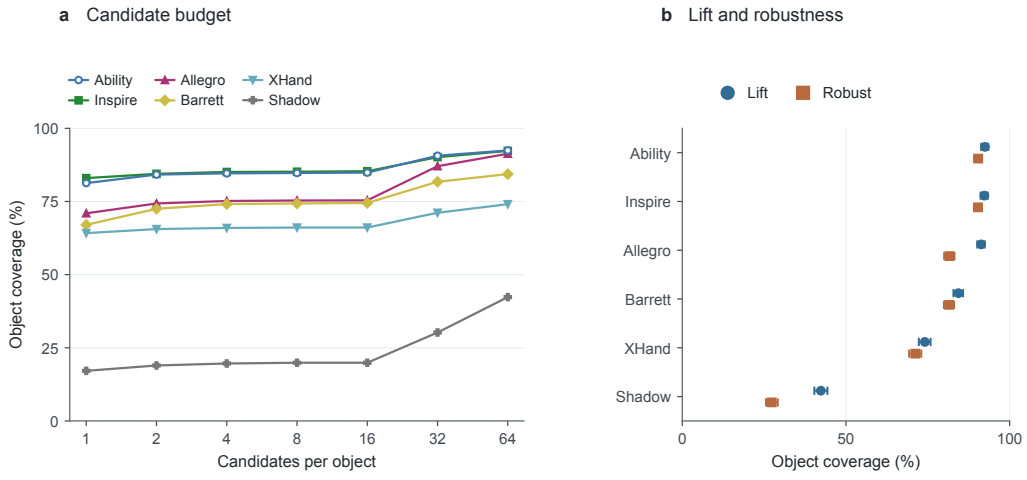


Figure S9: **Candidate budgets and execution outcomes.** Left: lift coverage at seven recorded budgets. Right: full-budget lift and robust coverage with Wilson 95% intervals, using the separate cohorts in Table S10. Lines connect recorded observations.

A.6.3 Robustness and closure displacement

Robustness is evaluated after the initial lift by combining holding with perturbations in mass, friction, and pose. Table S11 shows how many robust executions also satisfy the closure-translation bound, while Table S12 reports nominal, stronger, and combined perturbation outcomes on a separate stress-test cohort. The close correspondence between robust successes and bound-satisfying executions shows that the surviving candidates maintain stable object motion during closure and perturbation.

Table S11: Three-setting robust outcomes with an additional closure translation bound that measures object displacement during gripper closure.

Hand	N	Robust successes	Also bound	Also bound (%) ↑
Ability	2,367	2,140	2,140	90.41
Inspire	2,353	2,126	2,125	90.31
Allegro	2,372	1,936	1,935	81.58
Barrett	2,372	1,934	1,933	81.49
XHand	2,315	1,648	1,648	71.19
Shadow	2,372	648	646	27.23

Table S12: Separate stress-test cohort: nominal, stronger perturbation, and combined stress rates. N gives the number of evaluated objects for each hand.

Hand	N	Nominal (%) ↑	Stronger (%) ↑	Combined (%) ↑
Ability	2,372	91.6	91.1	90.6
Inspire	2,297	91.3	90.6	89.7
Allegro	2,372	86.3	80.7	80.4
Barrett	2,372	82.6	81.7	81.4
XHand	2,334	72.9	73.0	71.8
Shadow	2,353	40.8	40.0	38.7

The same candidate record also stores semantic-part hits, penetration, contact engagement, closure displacement, and physics outcomes, allowing the three experiments above to use one asset-side annotation format.

A.7 Policy Data, Training and Evaluation

A.7.1 Data budget and asset selection

For each of the four asset libraries, we train a separate policy on each of the five RoboTwin tasks using 100 successful demonstrations per task. This yields 20 policies trained on a total of 2,000 demonstrations, with the same demonstration budget across libraries.

Matching episode counts gives each source the same demonstration budget. The resulting comparison captures how the released asset coverage, trajectories, category match, and preparation choices combine to support policy learning. EmbodiedGen v2 uses one source asset per task, whereas the evaluation includes varied task objects, so the scores also reveal the value of broader task coverage.

A.7.2 Training and evaluation settings

Training. All four libraries use 100 demonstrations per task to fine-tune the $\pi_{0.5}$ base model with training seed 0 and global batch size 32. Evaluation uses the final checkpoint at step 19,999 for every source–task pair. Each source–task pair supplies its own training data and normalization statistics, and evaluation uses the final checkpoint without test-set selection.

Evaluation. Each of the 20 policies is evaluated on 100 clean and 100 randomized expert-solvable scenes, for 4,000 episodes. Clean and randomized suites are sampled independently. Randomization jointly varies visual context, table height, lighting, and instructions. Task success is the native RoboTwin success predicate; timeouts, planning failures, and infrastructure errors remain failures in each fixed denominator of 100. Each library has ten infrastructure-error episodes, all in the randomized condition.

The evaluation uses frozen suites. All libraries are evaluated using the same 100 seeds for each task–condition pair.

A.7.3 AGGREGATION AND TASK-LEVEL RESULTS

Macro averages weight the five tasks equally. The strongest baseline is ManiTwin in both modes: 55.4% clean and 31.2% randomized, compared with 71.2% and 41.8% for RFW-R. RFW-R leads on all five clean tasks and four randomized tasks; on randomized Move Pillbottle Pad, PhysX-3D reaches 9% and RFW-R 6%.

A.7.4 TASK-SEMANTIC CORRESPONDENCE BETWEEN SIMULATION AND REAL SETUPS

Figure S10 complements the paired simulation and real-robot views in Section 4.4 with temporal sequences from five recorded task families. The frames illustrate the progression of object manipulation through grasp, lift, and place stages.

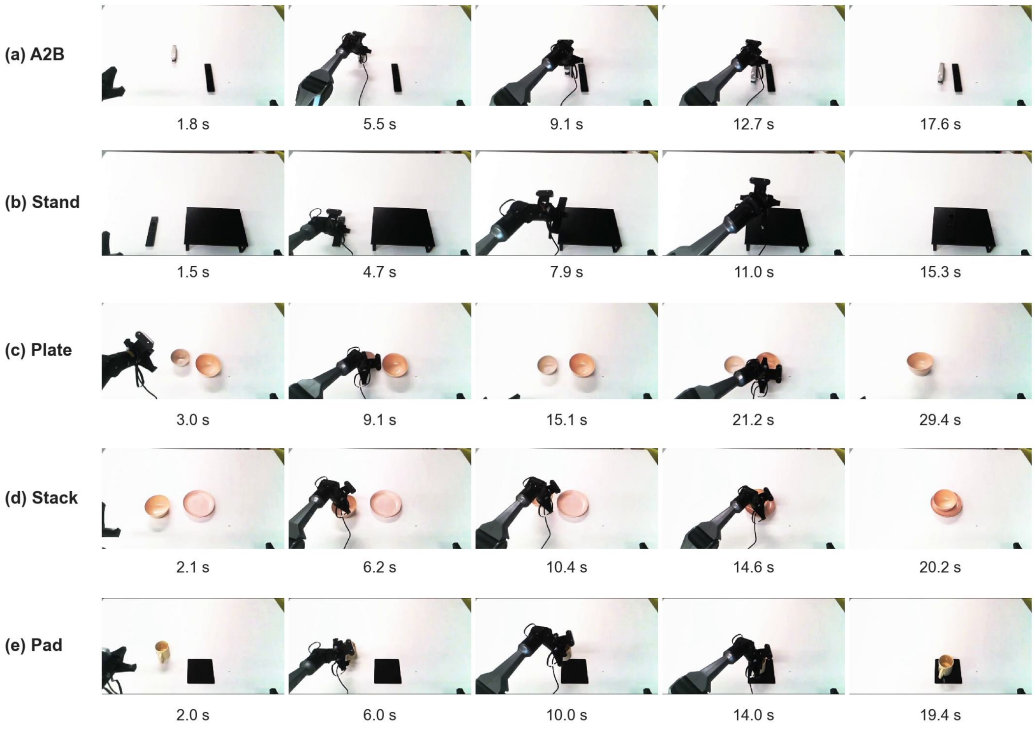


Figure S10: **Selected real-robot rollout sequences.** Rows show (a) place A left of B, (b) place on stand, (c) place on plate, (d) stack bowls, and (e) place on pad. Each row contains frames at 10%, 30%, 50%, 70%, and 97% of its recorded video. The pad example uses a real mug.