From Seeing to Simulating: Generative High-Fidelity Simulation with Digital Cousins for Generalizable Robot Learning and Evaluation

Jasper Lu[1*], **Zhenhao Shen**[1*], **Yuanfei Wang**[1*], **Shugao Liu**[1], **Shengqiang Xu**[1], **Shawn Xie**[1], **Jingkai Xu**[1], **Feng Jiang**[1], **Jade Yang**[1], **Ruihai Wu**[1†]

[1]Peking University

*Equal contribution, †Corresponding author Contact email: wuruihai@pku.edu.cn

Webpage: https://stubborn111.github.io/WorldComposer/

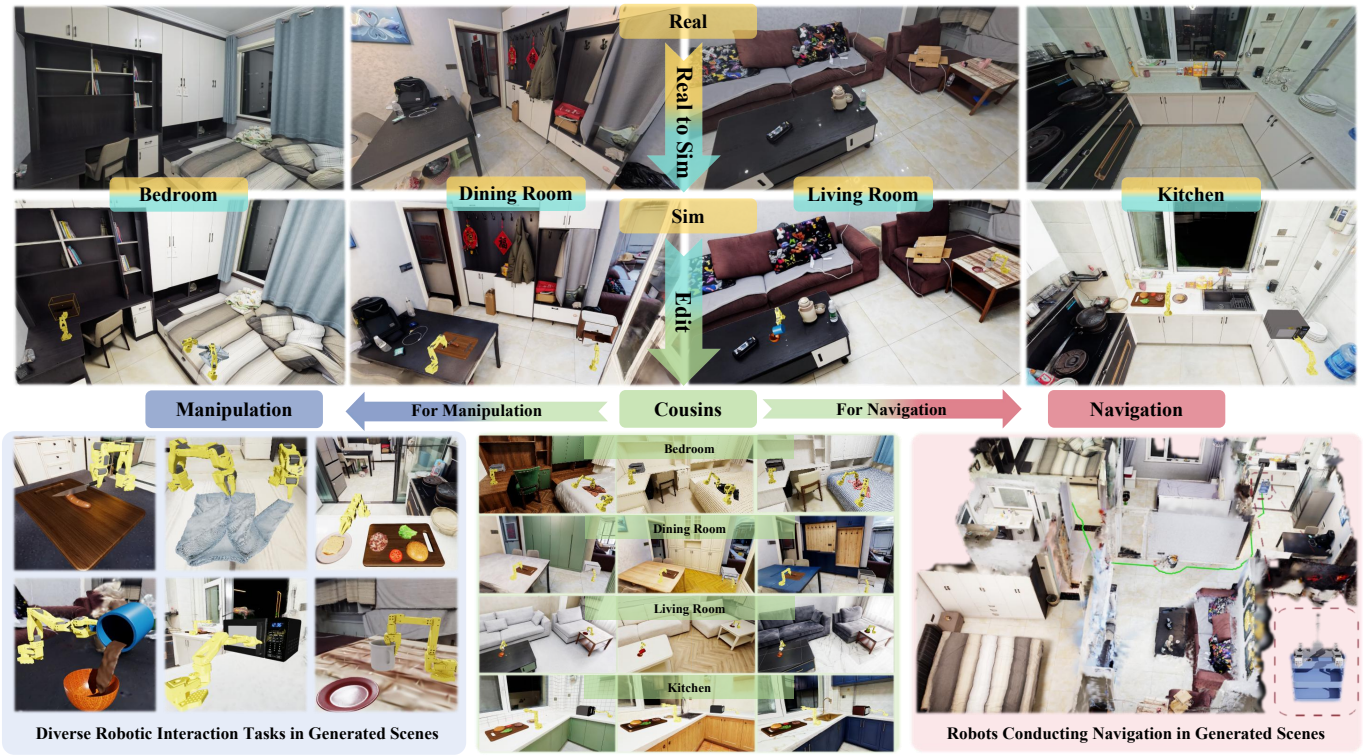

Fig. 1: **WorldComposer** generates high-fidelity simulation scenes from real-world panorama, and further generates diverse cousin scenes through editing. The generated rooms can be seamlessly stitched together into multi-room environments for navigation. Combined with interactive tasks, we provide a platform for generalizable learning and evaluation.

***Abstract*—Learning robust robot policies in real-world environments requires diverse data augmentation, yet scaling real-world data collection is costly due to the need for acquiring physical assets and reconfiguring environments. Therefore, augmenting real-world scenes into simulation has become a practical augmentation for efficient learning and evaluation. We present a generative framework that establishes a generative real-to-sim mapping from real-world panoramas to high-fidelity simulation scenes, and further synthesize diverse cousin scenes via semantic and geometric editing. Combined with high-quality physics engines and realistic assets, the generated scenes support interactive manipulation tasks. Additionally, we incorporate multi-room stitching to construct consistent large-scale environments for long-horizon navigation across complex layouts. Experiments demonstrate a strong sim-to-real correlation validating our platform's fidelity, and show that extensively scaling up data generation leads to significantly better generalization to unseen scene and object variations, demonstrating the effectiveness of Digital Cousins for generalizable robot learning and evaluation.**

I. Introduction

Robust robot learning in real-world environments depends heavily on access to diverse and realistic training data [18, 2, 4]. In particular, manipulation policies are highly sensitive to variations in scene layout, object geometry, appearance, and physical interactions. Collecting such diverse data directly in the real world, however, remains costly and labor-intensive, as it typically requires acquiring additional physical assets, re-

peatedly reconfiguring environments, and conducting extensive trials under varied conditions. These challenges significantly affect the policy robustness and generalization.

Simulation provides a practical augmentation by enabling efficient data generation and controlled experimentation. With modern physics engines [37, 14] and increasingly realistic assets [30, 6], simulation has become a valuable tool for robot learning. Nevertheless, constructing simulation environments that faithfully reflect real-world scenes remains challenging. Existing approaches often rely on manual scene modeling, predefined templates, or limited domain randomization. While effective in constrained settings, these methods either require substantial human effort or fail to capture the structural and semantic complexity of real-world environments.

Recent advances [26, 48, 22] in generative models offer new opportunities to bridge the gap between real-world perception and simulation. By leveraging visual observations, it becomes possible to automatically reconstruct simulation environments from real-world data. However, most prior work focuses on generating a single digital replica of a scene, providing limited support for systematic variation and data augmentation. As a result, the generated environments often lack the diversity needed to train and evaluate policies that generalize across unseen scene and object variations.

In this paper, as illustrated in Fig. 1, we present a generative framework, **WorldComposer**, that constructs high-fidelity simulation scenes directly from real-world panoramas. Leveraging generative models, our framework establishes an efficient real-to-sim mapping that produces realistic simulation environments without manual scene modeling. The framework supports multi-room scene construction, enabling the composition of complex environments suitable for scene-level tasks like navigation. Crucially, our framework is designed to operate in conjunction with high-quality physics engines [37, 30] and realistic asset libraries that span a wide range of object categories and interaction types. This combination enables interactive manipulation tasks beyond static scene generation.

Moreover, our framework supports efficient generation of diverse digital cousins of both scenes and objects through semantic and geometric editing. These digital cousins introduce controlled variations in layout, geometry, and object configurations, significantly expanding the diversity of interactive experiences available for training. By augmenting both environments and assets, our approach provides rich data for improving policy robustness and generalization.

The generated environments form a platform for generalizable robot learning and evaluation. By training and evaluating manipulation policies across diverse digital cousins, we enable systematic assessment of policy generalization under controlled scene and object variations. Extensive experiments demonstrate that policies trained on our platform generalize better to unseen scenes and objects. Moreover, we observe a strong correlation between policy performance in simulation and in the real world, indicating that the proposed platform provides a reliable basis for both learning and evaluation.

In summary, our main contributions are as follows:

- **Generative Real-to-Sim Scene Construction:** We present a generative framework that constructs high-fidelity multi-room simulation scenes directly from real-world panoramas, enabling an automated real-to-sim mapping without manual scene modeling.
- **High-Fidelity Interactive Simulation with Digital Cousins:** We enable efficient generation of diverse digital cousins of real-world scenes and objects through editing, resulting in interactive simulation environments with high visual and physical fidelity.
- **Generalizable Robot Learning and Evaluation Platform:** Extensive experiments show that policies trained on our platform achieve improved generalization across unseen scenes and objects. Besides, results in simulation exhibit strong correlation with real-world performance.

II. Related Work

A. Multimodal World Models

Recent advancements in world model have been propelled by large-scale video generation models [25, 39, 40, 49, 38, 5] function as powerful "latent simulators," yet inherently lack the explicit 3D geometry and physics required for precise robotic planning.Therefore, emerging works attempt to lift 2D generation into 3D space [24, 43, 48, 55]. And generative 3D methods like Marble [26] directly synthesize consistent, editable environments. Leveraging Marble for scalable scene generation, we bridge the interactivity gap by populating these environments with physics-enabled assets to construct fully interactive "Digital Cousins."

B. High-Fidelity Simulation Platforms

To facilitate sim-to-real transfer, simulation environments have advanced in **visual** and **physical** fidelity. For visual realism, platforms have shifted from procedural environments [20, 32, 47, 27] to photo-realistic scenes using digital twins [35, 6] and advanced rendering [36, 28, 13, 14]. Simultaneously, physical fidelity has evolved beyond rigid-body dynamics to model fluids [52], thin-shells [50, 45, 33] and volumetric objects [17, 16]. Our pipeline synergizes these frontiers, combining 3DGS-based reconstruction with rigorous hybrid physics to generate interactive and high-fidelity environments.

C. Real-to-Sim Environment Creation

The field of Real-to-Sim has evolved from procedural layout generation [11] to photorealistic digitization powered by 3D Gaussian Splatting (3DGS) [23]. To enable rich contact interactions, a surge of recent studies has integrated 3DGS with physics engines, creating closed-loop simulators that function as effective "Digital Twins" for robotic manipulation [42, 22, 29, 21, 15]. Concurrently, these high-fidelity environments are increasingly leveraged as scalable benchmarks for policy evaluation, demonstrating strong correlations between simulation metrics and real-world performance [19, 1, 10, 41]. Our work extends this trajectory by introducing a generative pipeline that expands static twins into diverse "Digital Cousins," serving as both a scalable engine for data augmentation and a rigorous testbed for generalization evaluation.

Fig. 2: **Overview of WorldComposer Environment Generation.** Our framework enables the rapid creation of interactive, high-fidelity simulation environments from real-world data. (Top) **Generative Real-to-Sim Scene Construction**: Using panoramic captures and our Marble engine, we reconstruct a "Sim Twin" and generate diverse "Sim Cousins" through prompt-based world editing. (Middle) **Multi-Room Stitching**: To handle complex environments, we use panorama matching and ICP refinement to stitch different rooms. (Bottom) **Asset Library & Generalizable Robot Learning**: We maintain high-fidelity simulation engine and a library of diverse assets. This culminates in a platform for generalizable robot learning and evaluation.

III. Generative High-Fidelity Environments with Digital Cousins

As illustrated in Fig. 2, our proposed framework **WorldComposer** generates high-fidelity, interactive simulation environments from real-world scenarios. The framework consists of three core stages: (1) **Automated Real-to-Sim Scene Generation**, where single-view panoramas are transformed into editable 3D Gaussian Splatting and Collision Mesh (both Digital Twins and Cousins); (2) **Multi-Room Stitching**, where individual generated rooms are spatially aligned and merged to form large-scale, navigable environments; and (3) **High-Fidelity Interactive Simulation**, where static scenes are enhanced with high-fidelity physics solvers, logic-aware interaction mechanisms, and high-quality assets. With the generated interactive environments with digital cousins, we can further generalize the robot policy learning and evaluation.

A. Generative Real-to-Sim Scene Construction

The foundation of our framework is to convert simple visual inputs into interactable environments. We adopt a coarse-to-fine generation strategy, progressing from panoramas to static "Digital Twins", and then to diverse "Digital Cousins".

From Panorama to Digital Twin. Given a single 360-degree panorama collected from the real world, we leverage Marble [26], a multimodal World Model, to reconstruct the scene geometry and texture. As shown in the "World Gen-

eration" module of Fig. 2, this process produces two key components: (i) Visual Rendering, a set of 3D Gaussian Splats that enables photorealistic rendering while capturing complex lighting and material effects; and (ii) Collision Mesh, a corresponding mesh that supports robot-environment interaction by providing rigid bodies for physics simulation.

To integrate these components into the Isaac Sim simulator, we utilize *3DGRUT* to convert the Gaussian PLY data and collision meshes into the Universal Scene Description (USD) format, ensuring compatibility with the Omniverse ecosystem.

Prompt-Driven Editing for Digital Cousins. A key limitation of simple real-to-sim Digital Twin scene generation is the lack of diversity, which hinders generalizable robot learning and evaluation. To address this, we leverage the editing capability of Marble. By conditioning the model on natural language prompts (e.g., "a kitchen with wooden textures" or "modern style layout"), we can modify the visual appearance and semantic layout of the reconstructed scene. This process generates Digital Cousins, environments that retain the structural logic of the original Digital Twin while introducing significant visual and layout variations. This one-to-many generation capability is crucial for domain randomization and generalizable policy learning.

B. Multi-Room Stitching

Since a 360-degree panorama typically covers only a single room, we need to stitch multiple room-level scenes into a house-scale, navigable environment. To this end, we introduce a robust multi-room stitching pipeline that spatially aligns discrete 3D Gaussian Splatting rooms generated by Marble into a cohesive floor, providing a comprehensive evaluation platform for long-horizon tasks.

1) Coarse Alignment via Panoramic Feature Matching: Stitching independent rooms is challenging due to the absence of global coordinates. While Marble synthesizes 3D representations from single 360-degree panoramas, generative 3DGS often exhibits viewpoint-dependent artifacts or geometric hallucination when departing from the sampling center, making traditional Structure-from-Motion (SfM) pipelines [44] unreliable for direct registration.

To ensure accuracy, we leverage overlapping visual cues between the original adjacent panoramas. Local features are extracted via **SuperPoint** [12] and matched using **LightGlue** [31]. We decompose the resulting essential matrix into a relative rotation R_{ab} and unit translation $\hat{t}_{ab}$. To resolve monocular scale ambiguity, we estimate the metric scale α by aligning the reconstructed ground plane with the known camera height h. Specifically, if $\mathcal{P}_g$ represents the set of triangulated ground points in the unit-scale coordinate system, the metric translation t_{ab} is recovered by determining the scale factor α as follows:

$$t_{ab} = \alpha \hat{t}_{ab}, \quad \text{where } \alpha = \frac{h}{\text{median}(\{\mathbf{n}^\top p_i \mid p_i \in \mathcal{P}_g\})} \tag{1}$$

where $\mathbf{n}$ is the floor plane normal. This yields a metric-aware coarse pose $T_{coarse} = [R_{ab}|t_{ab}]$ to initialize scene alignment.

2) Fine Refinement via Geometric ICP: To eliminate residual errors and ensure physical continuity, we perform geometric refinement on dense point clouds $\mathcal{P}_a, \mathcal{P}_b$ sampled from overlapping regions. We apply Point-to-Plane **Iterative Closest Point (ICP)** [7] to minimize the geometric distance:

$$E_{ICP} = \sum_i \left\| (T_{fine} \cdot p_a^i - p_b^j)^\top \mathbf{n}_b^j \right\|^2 \tag{2}$$

where T_{fine} is initialized by T_{coarse}. Upon convergence, the 3DGS kernels are merged into a unified coordinate system. Finally, the stitched 3DGS and collision meshes are converted into the USD format via *3DGRUT*, providing a seamless and navigable surface for embodied agents in Isaac Sim.

C. High-Fidelity Interactive Simulation

While the automated scene generation provides photorealistic backgrounds, realistic robotic interaction ultimately depends on manipulable objects. However, both the 3D Gaussian Splats and the derived collision meshes are static. To enable interactive simulation, we integrate the reconstructed scenes with high-quality assets supported by high-fidelity physics solvers. In particular, we curate a comprehensive asset library spanning three categories:

- **Rigid Objects:** We apply accurate convex decomposition to support stable grasping and reliable collision handling, providing a baseline for standard manipulation tasks.
- **Articulated Objects:** For objects with kinematic chains, we define precise joint limits and drive mechanisms. This enables realistic kinematic interactions, such as opening microwaves, closing drawers, while maintaining valid physical states throughout the interaction.
- **Deformable Objects:** Following the simulation engine in LeHome [30], we employ advanced physics solvers tailored to specific material properties to handle complex non-rigid dynamics. Specifically, we utilize **Position-Based Dynamics** for diverse garments; the **Finite Element Method** for elastic volumetric objects like soft foods; and **Dynamic Grid Method** to simulate fluids.

Scene Composition. The final step involves populating the generated Digital Cousins with these assets. We leverage an LLM to provide common-sense, semantically grounded placement priors (e.g., placing appliances in kitchens). We then estimate the pose of each target placement location and instantiate assets with physically valid alignment. Several representative simulation environments are shown in Fig. 3. This composition combines the visual realism of the reconstructed scene with the physical richness of our asset library, yielding a holistic environment for diverse embodied tasks.

D. Generalizable Robot Learning and Evaluation

Our proposed framework can serve as both a data engine for generalizable robot learning and a platform for generalization evaluation. It directly alleviates the scalability bottleneck of real-world data collection by transforming finite real-world captures into a continuously expandable stream of interactive simulation experiences.

Fig. 3: **Simulation Tasks and Digital Cousin Generation.** Our pipeline supports a diverse range of high-fidelity simulation tasks. And each row demonstrates the transition from a sparse real-world capture (**Real**) to a precise digital reconstruction (**Twin**), followed by the generation of multiple **Digital Cousins**.

Starting from a single panoramic image, we exploit the prompt-driven editing capabilities of Marble to synthesize a dense neighborhood of environment variants. We apply semantic and geometric edits to each scene, such as changes in texture, lighting, and layout. We also populate the environment with diverse interactive objects from our *high-fidelity asset library*. Together, these variations yield a broad and effectively unbounded training distribution. Training on large collections of automatically generated Digital Cousins promotes robust, semantic-aware behaviors, rather than overfitting to a small set of fixed configurations.

Beyond data augmentation, our framework supports a high-fidelity evaluation platform that reflects real-world deployment. Instead of reporting performance on a single static test environment, we evaluate policies across a hierarchical set of generalization tiers with progressively larger distribution shifts. This design disentangles different failure modes by probing robustness under changes in appearance, scene structure, object instances, and their combinations, spanning settings from seen environments to novel compositions of unseen scenes and unseen objects. As a result, the evaluation provides a systematic test for zero-shot generalization, yielding metrics that more reliably correlate with real-world evaluation.

IV. Experiments

We evaluate **WorldComposer** in both simulation and the real world. We benchmark several state-of-the-art policies on tasks spanning rigid, articulated, and deformable interactions, and study three questions: whether Digital Cousin data improves generalization, whether our simulation evaluation correlates with real-world results, and whether stitched multi-room environments support long-horizon navigation.

Fig. 4: **Real-World Setup.** The hardware configuration consists of two lerobot arms and diverse object instances.

A. Experimental Setup

Simulation Setting. Built on **NVIDIA Isaac Sim**, our simulation features a kinematic "Digital Twin" of the physical robot. We integrate the **LeRobot** interface directly into the simulation loop to enforce a shared control stack. This ensures strict consistency between the simulated "Digital Cousins" and the real world, minimizing implementation gaps.

Real-World Setting. Our real-world setup features a bi-manual Leader-Follower system utilizing two LeRobot SO-101 follower arms and diverse objects, as illustrated in Fig. 4. The observation is captured by a top-down RGB camera. We utilize the LeRobot framework for unified hardware control and data collection, ensuring standardized data formats for training.

Tasks. As shown in Fig. 3, we select a suite of distinct manipulation tasks that span the full spectrum of physical complexity defined in our asset library:

TABLE I: We compare the performance of Diffusion Policy trained under two data settings, on novel scenes and objects.

	Set Tableware		Pour Water		Open Microwave		Close Drawer	
	Scene	Object	Scene	Object	Scene	Object	Scene	Object
100 Original data	0.61	0.30	0.64	0.46	0.59	0.33	0.78	0.48
100 Original + 200 Cousin data	0.68	0.50	0.70	0.70	0.62	0.56	0.87	0.72
	Fold Cloth		Cut Sausage		Assemble Burger		Average	
	Scene	Object	Scene	Object	Scene	Object	Scene	Object
100 Original data	0.45	0.29	0.96	0.86	0.57	0.47	0.66	0.46
100 Original + 200 Cousin data	0.54	0.47	0.96	0.90	0.63	0.58	0.71	0.63

TABLE II: We compare the performance of π_0 trained under two data settings, on novel scenes and objects.

	Set Tableware		Open Microwave		Fold Cloth		Average	
	Scene	Object	Scene	Object	Scene	Object	Scene	Object
100 Original data	0.84	0.53	0.82	0.58	0.67	0.55	0.78	0.55
100 Original + 200 Cousin data	0.89	0.78	0.88	0.81	0.72	0.69	0.83	0.76

- *Rigid Body Manipulation (Set Tableware):* This task involves picking up diverse rigid objects—specifically cups, mugs, or bowls—and precisely placing them onto a designated target plate.
- *Articulated Object Interaction (Open Microwave & Close Drawer):* These tasks involve manipulating constrained mechanisms. Specifically, the robot must grasp the handle to open a microwave and push to close an open drawer.
- *Deformable, Fluid & Hybrid Interaction:* This category covers complex dynamics: *Fold Cloth* involves folding a garment; *Pour Water* requires pouring liquid into a cup without spillage; and *Food Preparation* involves slicing a sausage and stacking a patty.

Evaluated policies. To demonstrate our environment's superiority for generalizable robot learning and evaluation, we evaluate four state-of-the-art methods for manipulation:

- **Action Chunking Transformer (ACT) [54]**, a Transformer-based policy utilizing temporal ensembling for precise motion generation.
- **Diffusion Policy (DP) [9]**, a commonly adopted generative policy that models multi-modal action distributions via diffusion processes, offering superior stability.
- **SmolVLA [46]**, an efficient, open-weight Vision-Language-Action (VLA) model designed to benchmark the performance of lightweight semantic-aware policies.
- **π_0 [3]**, a cutting-edge generalist VLA, representative for large-scale pre-trained robotic foundation models.

Data Collection & Metrics. We collected expert demonstrations for manipulation tasks using a Leader-Follower teleoperation setup operating at a control frequency of 30Hz. Regarding training data, we collected 50 expert demonstrations for standard baseline evaluations in both simulation and real-world settings. For specific manipulation experiments, the exact training data compositions vary by settings and are detailed in their respective tables.

Evaluation Metrics. We conduct **100 trials** per configuration in simulation and **20 trials** in the real world for evaluation. As the metric, we report the **Success Rate (SR)** across four generalization levels: *Train*, *Unseen Scene*, *Unseen Object*, and *Unseen Scene & Object*.

B. Efficacy of Digital Cousin Data

We assess the efficacy of our framework by adding diverse generated Digital Cousin data for generalizable robotic manipulation policies. Fig. 5 demonstrates the augmentation and co-training pipeline. With the experiments and analysis, we aim to answer: *Does adding diverse Digital Cousin data improve policy robustness and generalization?*

1) Analysis of Simulation Experiments. We trained policies using a set of original single-scene data and compared them against policies co-trained with generated cousin data.

- **Diffusion Policy:** As shown in Table I, adding augmented data significantly boosts performance across all categories. Notably, in physically complex tasks like *Pour Water* and *Cut Sausage*, the success rates in "Scene & Object" generalization settings improved by a large margin. This indicates that the high-fidelity physical engine in our simulation helps the model learn robust dynamics of complex interactions.
- **VLA Models:** As shown in Table II, for foundation models like π_0, the augmentation also yields positive gains, particularly in the *Set Tableware* task, proving that our high-fidelity visual generation is beneficial to large-scale pre-trained robotic foundation model.

2) Real-Sim-Real Transfer. We validate the utility of our generated data by deploying the policy in the real world deployment (Fig. 5 and Table III).

For the intrinsic quality of our generated data, remarkably, the policy trained solely on 50 Sim (Twin) data achieves generalization similar to the policy trained on 50 Real data. This parity demonstrates that our framework is able to function as a high-fidelity proxy for real-world interaction.

Besides, we observe significant gains when supplementing real data with generated data. As shown in Table III and Fig. 5,

TABLE III: **Real-World Evaluation Results.** We compare the performance of DP trained on different combinations of data. "Twin" denotes simulation scenes and objects reconstructed from real images, while "Cousin" denotes generated variations.

	Set Tableware		Open Microwave		Fold Cloth		Average	
	Scene	Object	Scene	Object	Scene	Object	Scene	Object
50 Real world data	0.30	0.30	0.50	0.45	0.20	0.05	0.33	0.27
50 Sim (Twin) data	0.40	0.30	0.45	0.35	0.15	0.10	0.33	0.25
100 Real world data	0.30	0.40	0.50	0.45	0.30	0.20	0.37	0.35
50 Real world + 50 Sim (Twin) data	0.25	0.40	0.50	0.50	0.25	0.15	0.33	0.35
50 Sim (Twin) + 50 Sim (Cousin) data	0.45	0.35	0.60	0.50	0.25	0.20	0.43	0.35
50 Real world + 100 Sim (Twin + Cousin) data	0.55	0.50	0.75	0.65	0.40	0.35	0.57	0.50

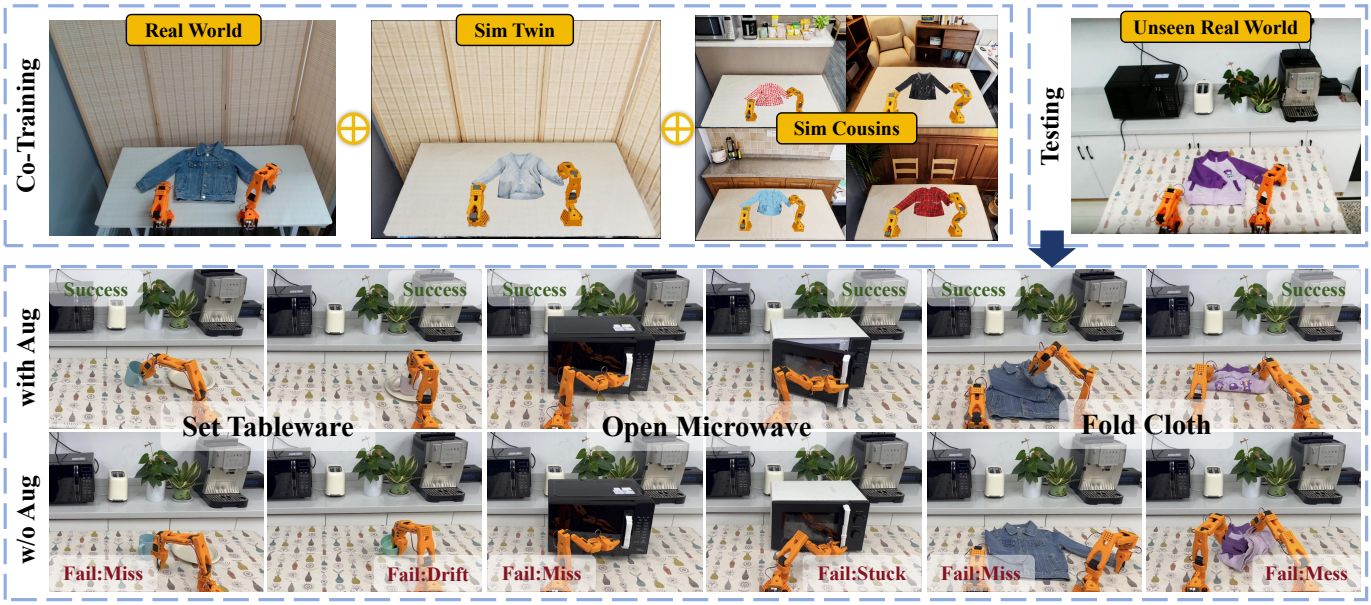

Fig. 5: **Real-to-Sim-to-Real Pipeline and Qualitative Analysis.** (Top) The co-training framework with real-to-sim cousins. (Bottom) The comparison of policy execution in unseen scenarios. Aug denotes the augmentation of generated data.

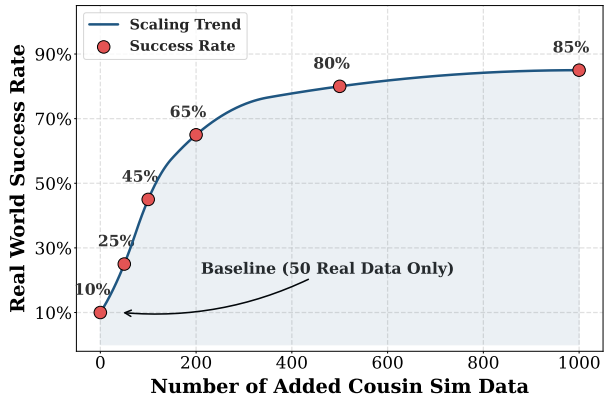

Fig. 6: **Performance with Increasing Cousin Sim Data.** Real-world success rate on Set Tableware task with increasing cousin sim data under the unseen scene and object variations.

the co-training configuration (50 Real + 100 Sim) outperforms the baseline (50 Real only) by a significant margin, nearly doubling success rates in generalization settings (0.57 vs. 0.33 in Unseen Scene). This confirms that "Digital Cousins" provide the critical environmental fidelity and diversity necessary to bridge the data scarcity gap.

Furthermore, we incrementally add up to 1,000 Digital Cousin trajectories to the 50 Real trajectories, on the most challenging *unseen scene & object* setting. The results in Fig. 6 demonstrate a consistent upward trend, with the success rate rising from **10%** to **85%**. The results demonstrates that large-scale data with variation generated by **WorldComposer** can effectively boost generalization of robotic manipulation policy.

C. Real-to-Sim Evaluation Correlation

A critical prerequisite for a reliable simulation evaluation platform is its ability to precisely reflect real-world performance. To validate the fidelity of our generated Digital Cousin environments, we conducted a rigorous correlation analysis by comparing success rates across three representative tasks: Set Tableware, Open Microwave, and Fold Cloth.

As illustrated in Fig. 7, we observe a strong positive linear relationship between the simulation and real-world success rates. The data points, representing various baseline methods (ACT, DP, SmolVLA, and π_0) across multiple generalization levels, cluster tightly around the identity line. The **Pearson**

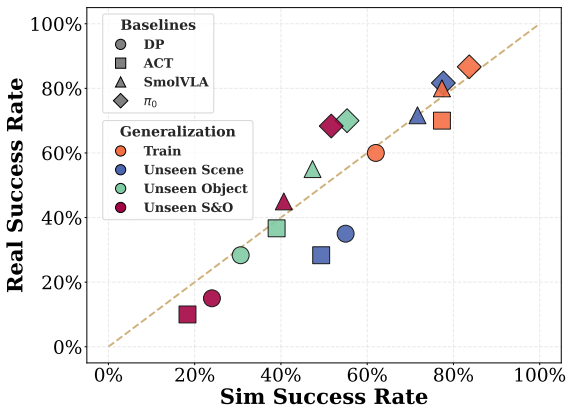

Fig. 7: **Sim-Real Evaluation Correlation.** The scatter plot illustrates the average success rates across three tasks (Set Tableware, Open Microwave, Fold Cloth) for various baselines (ACT, DP, SmolVLA, π_0) under four generalization levels.

correlation coefficient yields a high value of $r = 0.91$, which provides several key validations for our benchmark:

- **Real-Sim Alignment:** The high r-value demonstrates that our simulation evaluation aligns with the real world. A policy failure in our Digital Cousins, caused by collision or dynamic mismanagement, serves as a reliable predictor of failure in the real world.
- **Performance Ranking Preservation:** The differences of different architectures is consistently preserved across domains. For instance, the performance difference between VLA-based models (π_0, SmolVLA) and standard imitation policies (ACT, DP) can be mirrored.
- **Generalization Consistency:** The simulation successfully captures the difficulty gradient associated with unseen factors. The performance degradation observed in "Unseen Scene & Object" configurations in simulation is strongly correlated with the results obtained in corresponding real-world deployments.

This result establishes our system as a high-fidelity, reliable testbed for evaluating generalist policies at scale prior to real-world deployment.

D. Navigation in Generated Multi-Room Environments

To validate that the stitched environments are suitable for scene-level robotic tasks, we conduct **Zero-Shot Object-Goal Navigation (ZSON)** experiments using the movable XLeRobot. This task requires the robot to traverse complex, house-scale floorplans to locate interactive assets within a unified coordinate system.

- **Setup**: Five distinct interactive assets from our library are procedurally injected into the stitched layout. We conduct 20 evaluation episodes per asset (100 total), with starting positions sampled from rooms distant to the goal to necessitate long-horizon, cross-room planning.
- **Evaluated Method**: We employ **VLFM** [53], a zero-shot policy noted for its deployment efficiency. It utilizes a pre-trained VLM to extract semantic cues from RGB streams, identifying promising frontiers in unfamiliar environments without prior training.

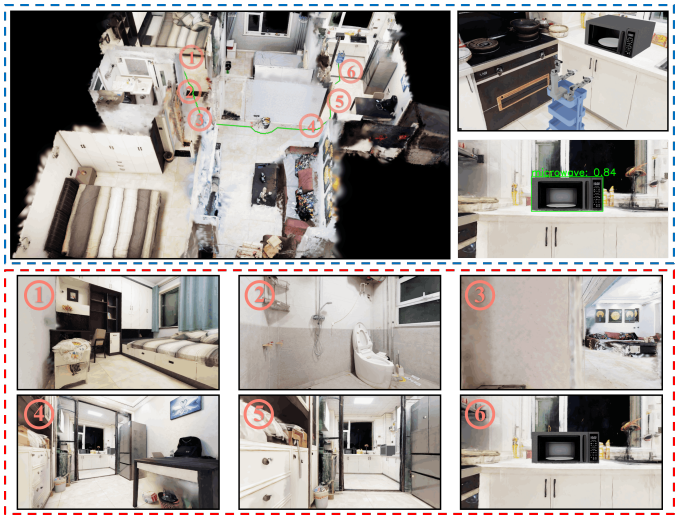

Fig. 8: **Multi-Room Navigation.** (Top) The navigation trajectory in the multi-room environment, and the goal state. (Bottom) First-person observations during navigation.

TABLE IV: **Navigation Results.** We report success rate of VLFM in stitched multi-room environments.

Task (Target Object)	SR ↑	SPL ↑
Microwave	0.70	0.58
Chair	0.60	0.47
Toilet	0.80	0.66
Oven	0.55	0.42
Refrigerator	0.75	0.61
Total Average	**0.68**	**0.55**

As detailed in Table IV, VLFM achieves a 68% average success rate, demonstrating that our stitched multi-room environments provide a physically robust and semantically consistent testbed, effectively supporting the validation of semantic reasoning and long-horizon navigation.

Fig. 8 provides visualization of navigation in stitched multi-room environments. The top panel illustrates a representative cross-room trajectory within the integrated floorplan, serving as an example of long-horizon movement in the stitched environment. The bottom panel displays a temporal sequence of first-person view (FPV) observations, highlighting the visual consistency and environmental fidelity during traversal.

V. Conclusion

We presented a generative framework that generates high-fidelity, interactive simulations with diverse digital cousins from real-world panoramas. Improved policy robustness and strong real-sim correlation validate the platform for generalizable robot learning and evaluation.

Future Work. While Marble generates the whole mesh of the scene, we will further explore instance-level decomposition, leveraging 3D semantic segmentation to replace static elements. Additionally, to address texture discontinuities at room junctions, we plan to investigate cross-scene radiance field fusion and end-to-end 3DGS optimization.

References

[1] Jad Abou-Chakra, Lingfeng Sun, Krishan Rana, Brandon May, Karl Schmeckpeper, Niko Suenderhauf, Maria Vittoria Minniti, and Laura Herlant. Real-is-sim: Bridging the sim-to-real gap with a dynamic digital twin, 2025. URL https://arxiv.org/abs/2504.03597.

[2] Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

[3] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. pi_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

[4] Chilam Cheang, Sijin Chen, Zhongren Cui, Yingdong Hu, Liqun Huang, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Xiao Ma, Hao Niu, Wenxuan Ou, Wanli Peng, Zeyu Ren, Haixin Shi, Jiawen Tian, Hongtao Wu, Xin Xiao, Yuyang Xiao, Jiafeng Xu, and Yichu Yang. Gr-3 technical report, 2025. URL https://arxiv.org/abs/2507.15493.

[5] Boyuan Chen, Tianyuan Zhang, Haoran Geng, Kiwhan Song, Caiyi Zhang, Peihao Li, William T. Freeman, Jitendra Malik, Pieter Abbeel, Russ Tedrake, Vincent Sitzmann, and Yilun Du. Large video planner enables generalizable robot control, 2025. URL https://arxiv.org/abs/2512.15840.

[6] Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Qiwei Liang, Zixuan Li, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025.

[7] Yang Chen and Gérard Medioni. Object modelling by registration of multiple range images. *Image and Vision Computing*, 10(3):145–155, 1992. ISSN 0262-8856. doi: https://doi.org/10.1016/0262-8856(92)90066-C. URL https://www.sciencedirect.com/science/article/pii/026288569290066C. Range Image Understanding.

[8] Zoey Chen, Aaron Walsman, Marius Memmel, Kaichun Mo, Alex Fang, Karthikeya Vemuri, Alan Wu, Dieter Fox, and Abhishek Gupta. Urdformer: A pipeline for constructing articulated simulation environments from real-world images, 2024. URL https://arxiv.org/abs/2405.11656.

[9] 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. *The International Journal of Robotics Research*, page 02783649241273668, 2023.

[10] Prithwish Dan, Kushal Kedia, Angela Chao, Edward Weiyi Duan, Maximus Adrian Pace, Wei-Chiu Ma, and Sanjiban Choudhury. X-sim: Cross-embodiment learning via real-to-sim-to-real, 2025. URL https://arxiv.org/abs/2505.07096.

[11] Matt Deitke, Rose Hendrix, Luca Weihs, Ali Farhadi, Kiana Ehsani, and Aniruddha Kembhavi. Phone2proc: Bringing robust robots into our chaotic world, 2022. URL https://arxiv.org/abs/2212.04819.

[12] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description, 2018. URL https://arxiv.org/abs/1712.07629.

[13] Haoran Geng, Feishi Wang, Songlin Wei, Yuyang Li, Bangjun Wang, Boshi An, Charlie Tianyue Cheng, Haozhe Lou, Peihao Li, Yen-Jen Wang, et al. Roboverse: Towards a unified platform, dataset and benchmark for scalable and generalizable robot learning. *arXiv preprint arXiv:2504.18904*, 2025.

[14] Xiangyu Guo, Zhanqian Wu, Kaixin Xiong, Ziyang Xu, Lijun Zhou, Gangwei Xu, Shaoqing Xu, Haiyang Sun, Bing Wang, Guang Chen, Hangjun Ye, Wenyu Liu, and Xinggang Wang. Genesis: Multimodal driving scene generation with spatio-temporal and cross-modal consistency, 2025. URL https://arxiv.org/abs/2506.07497.

[15] Xiaoshen Han, Minghuan Liu, Yilun Chen, Junqiu Yu, Xiaoyang Lyu, Yang Tian, Bolun Wang, Weinan Zhang, and Jiangmiao Pang. Re^{3}sim: Generating high-fidelity simulation data via 3d-photorealistic real-to-sim for robotic manipulation, 2025. URL https://arxiv.org/abs/2502.08645.

[16] Eric Heiden, Miles Macklin, Yashraj Narang, Dieter Fox, Animesh Garg, and Fabio Ramos. Disect: A differentiable simulation engine for autonomous robotic cutting. *arXiv preprint arXiv:2105.12244*, 2021.

[17] Zhiao Huang, Yuanming Hu, Tao Du, Siyuan Zhou, Hao Su, Joshua B Tenenbaum, and Chuang Gan. Plasticinelab: A soft-body manipulation benchmark with differentiable physics. *arXiv preprint arXiv:2104.03311*, 2021.

[18] 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.

[19] 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, 2025. URL https://arxiv.org/

abs/2512.16881.

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

[21] Yufei Jia, Guangyu Wang, Yuhang Dong, Junzhe Wu, Yupei Zeng, Haonan Lin, Zifan Wang, Haizhou Ge, Weibin Gu, Kairui Ding, Zike Yan, Yunjie Cheng, Yue Li, Ziming Wang, Chuxuan Li, Wei Sui, Lu Shi, Guanzhong Tian, Ruqi Huang, and Guyue Zhou. Discoverse: Efficient robot simulation in complex high-fidelity environments, 2025. URL https://arxiv.org/abs/2507.21981.

[22] Guangqi Jiang, Haoran Chang, Ri-Zhao Qiu, Yutong Liang, Mazeyu Ji, Jiyue Zhu, Zhao Dong, Xueyan Zou, and Xiaolong Wang. Gsworld: Closed-loop photorealistic simulation suite for robotic manipulation, 2025. URL https://arxiv.org/abs/2510.20813.

[23] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering, 2023. URL https://arxiv.org/abs/2308.04079.

[24] Geonung Kim, Janghyeok Han, and Sunghyun Cho. Videofrom3d: 3d scene video generation via complementary image and video diffusion models, 2025. URL https://arxiv.org/abs/2509.17985.

[25] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models, 2025. URL https://arxiv.org/abs/2412.03603.

[26] World Labs. Marble, 2025. URL https://marble.worldlabs.ai. Accessed: 2026-01-25.

[27] Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, et al. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. *arXiv preprint arXiv:2108.03272*, 2021.

[28] Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *Conference on Robot Learning*, pages 80–93. PMLR, 2023.

[29] Xinhai Li, Jialin Li, Ziheng Zhang, Rui Zhang, Fan Jia, Tiancai Wang, Haoqiang Fan, Kuo-Kun Tseng, and Ruiping Wang. Robogsim: A real2sim2real robotic gaussian splatting simulator, 2025. URL https://arxiv.org/abs/2411.11839.

[30] Zeyi Li, Jade Yang, Jingkai Xu, Shangbin Xie, Yuran Wang, Zhenhao Shen, Tianxing Chen, Yan Shen, Wenjun Li, Yukun Zheng, Chaorui Zhang, Ming Chen, Chen Xie, and Ruihai Wu. Lehome: A simulation environment for deformable object manipulation in household scenarios. In *IROS 2025 - 5th Workshop on RObotic MAnipulation of Deformable Objects: holistic approaches and challenges forward*, 2025. URL https://openreview.net/forum?id=rEDd1HorJl.

[31] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local Feature Matching at Light Speed. In *ICCV*, 2023.

[32] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *NeurIPS*, 2023.

[33] Haoran Lu, Ruihai Wu, Yitong Li, Sijie Li, Ziyu Zhu, Chuanruo Ning, Yan Zhao, Longzan Luo, Yuanpei Chen, and Hao Dong. Garmentlab: A unified simulation and benchmark for garment manipulation. *NeurIPS*, 2024.

[34] Abhiram Maddukuri, Zhenyu Jiang, Lawrence Yunliang Chen, Soroush Nasiriany, Yuqi Xie, Yu Fang, Wenqi Huang, Zu Wang, Zhenjia Xu, Nikita Chernyadev, Scott Reed, Ken Goldberg, Ajay Mandlekar, Linxi Jim Fan, and Yuke Zhu. Sim-and-real co-training: A simple recipe for vision-based robotic manipulation. *ArXiv*, abs/2503.24361, 2025. URL https://api.semanticscholar.org/CorpusID:277467951.

[35] Yao Mu, Tianxing Chen, Zanxin Chen, Shijia Peng, Zhiqian Lan, Zeyu Gao, Zhixuan Liang, Qiaojun Yu, Yude Zou, Mingkun Xu, Lunkai Lin, Zhiqiang Xie, Mingyu Ding, and Ping Luo. Robotwin: Dual-arm robot benchmark with generative digital twins. In *CVPR*, 2025.

[36] 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. *arXiv preprint arXiv:2406.02523*, 2024.

[37] NVIDIA. Isaac Sim. URL https://github.com/isaac-sim/IsaacSim.

[38] NVIDIA, :, Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, Prithvijit Chattopadhyay, Mike Chen, Yongxin Chen, Yu Chen, Shuai Cheng, Yin Cui, Jenna Diamond, Yifan Ding, Jiaojiao Fan, Linxi Fan, Liang Feng, Francesco Ferroni, Sanja Fidler, Xiao Fu, Ruiyuan Gao, Yunhao Ge, Jinwei Gu, Aryaman Gupta, Siddharth Gururani, Imad El Hanafi, Ali Hassani, Zekun Hao, Jacob Huffman, Joel Jang, Pooya Jannaty, Jan Kautz, Grace Lam, Xuan Li, Zhaoshuo Li, Maosheng Liao, Chen-Hsuan Lin, Tsung-Yi Lin, Yen-Chen Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Yifan Lu, Alice

Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Seungjun Nah, Yashraj Narang, Abhijeet Panaskar, Lindsey Pavao, Trung Pham, Morteza Ramezanali, Fitsum Reda, Scott Reed, Xuanchi Ren, Haonan Shao, Yue Shen, Stella Shi, Shuran Song, Bartosz Stefaniak, Shangkun Sun, Shitao Tang, Sameena Tasmeen, Lyne Tchapmi, Wei-Cheng Tseng, Jibin Varghese, Andrew Z. Wang, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Jiashu Xu, Dinghao Yang, Xiaodong Yang, Haotian Ye, Seonghyeon Ye, Xiaohui Zeng, Jing Zhang, Qinsheng Zhang, Kaiwen Zheng, Andrew Zhu, and Yuke Zhu. World simulation with video foundation models for physical ai, 2025. URL https://arxiv.org/abs/2511.00062.

[39] OpenAI. Sora, 2024. URL https://openai.com/sora/. Accessed: 2026-01-25.

[40] OpenAI. Sora 2, 2025. URL https://openai.com/zh-Hans-CN/index/sora-2/. Accessed: 2026-01-25.

[41] Nicholas Pfaff, Evelyn Fu, Jeremy Binagia, Phillip Isola, and Russ Tedrake. Scalable real2sim: Physics-aware asset generation via robotic pick-and-place setups, 2025. URL https://arxiv.org/abs/2503.00370.

[42] Mohammad Nomaan Qureshi, Sparsh Garg, Francisco Yandun, David Held, George Kantor, and Abhisesh Silwal. Splatsim: Zero-shot sim2real transfer of rgb manipulation policies using gaussian splatting, 2024. URL https://arxiv.org/abs/2409.10161.

[43] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control, 2025. URL https://arxiv.org/abs/2503.03751.

[44] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[45] Daniel Seita, Pete Florence, Jonathan Tompson, Erwin Coumans, Vikas Sindhwani, Ken Goldberg, and Andy Zeng. Learning to Rearrange Deformable Cables, Fabrics, and Bags with Goal-Conditioned Transporter Networks. In *ICRA*, 2021.

[46] Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, et al. Smolvla: A vision-language-action model for affordable and efficient robotics. *arXiv preprint arXiv:2506.01844*, 2025.

[47] Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. Habitat 2.0: Training home assistants to rearrange their habitat. *NeurIPS*, 2021.

[48] GigaWorld Team, Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Haoyun Li, Jiagang Zhu, Kerui Li, Mengyuan Xu, Qiuping Deng, Siting Wang, Wenkang Qin, Xinze Chen, Xiaofeng Wang, Yankai Wang, Yu Cao, Yifan Chang, Yuan Xu, Yun Ye, Yang Wang, Yukun Zhou, Zhengyuan Zhang, Zhehao Dong, and Zheng Zhu. Gigaworld-0: World models as data engine to empower embodied ai, 2025. URL https://arxiv.org/abs/2511.19861.

[49] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

[50] Yian Wang, Juntian Zheng, Zhehuan Chen, Zhou Xian, Gu Zhang, Chao Liu, and Chuang Gan. Thin-shell object manipulations with differentiable physics simulations. In *ICLR*, 2023.

[51] 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, 2023.

[52] Zhou Xian, Bo Zhu, Zhenjia Xu, Hsiao-Yu Tung, Antonio Torralba, Katerina Fragkiadaki, and Chuang Gan. Fluidlab: A differentiable environment for benchmarking complex fluid manipulation. *arXiv preprint arXiv:2303.02346*, 2023.

[53] Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. Vlfm: Vision-language frontier maps for zero-shot semantic navigation. In *International Conference on Robotics and Automation (ICRA)*, 2024.

[54] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

[55] Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. Tesseract: Learning 4d embodied world models, 2025. URL https://arxiv.org/abs/2504.20995.

APPENDIX

A. Technical Implementation Details

Real-to-Sim Alignment and Physical Fidelity. To anchor assets and ensure stable contacts within the manipulation workspace, we extract the tabletop via SAM and RANSAC, flattening the collision mesh onto the fitted plane ($\leq$ 1cm variance). To validate the sim-to-real object gap, we propped up the same garment's center using an identical cylinder in both simulation and the real world, which yielded only a 6% bounding box shrinkage discrepancy, confirming our precise physical alignment.

Implementation of Fluid Interaction (Pour Water). For complex physical interactions such as the Pour Water task, we simulate fluid-solid dynamics via Position-Based Dynamics (PBD) and render continuous liquid surfaces using Isosurfaces. The water volume is initialized by uniformly sampling particles inside the container's convex hull. A trial is considered successful if $> 60\%$ of the fluid particles enter the target receptacle's bounding box.

Multi-Room Stitching and Navigation. We sample portal panoramas to provide visual anchors for robust stitching, enabling multi-view fusion where overlapping regions compensate for artifacts. Our 1:1 real-world navigation experiments confirm high sim-to-real behavioral alignment (Table V). Beyond the success rate, minimal Dynamic Time Warping (DTW) distances and consistent Value Map topologies (Fig. 9) further validate the semantic and geometric fidelity required for reliable zero-shot evaluation.

Digital Cousins Generation and Auto-Prompting. Using predefined templates, an LLM automatically generates diverse prompts for scene editing. Digital Cousins feature two core augmentations: Scene Cousins (Marble-edited backgrounds) and Object Cousins (assets with varied geometries and physics), supplemented by lighting, pose, and texture randomizations. Each task contains 5 scene variants and 10 object variants.

Automated Data Collection and Evaluation. To drive data collection, our automated engine deploys 14 parameterized scripted skills. During real-world evaluation, Twin data isolates environmental alignment via kinematic playback (eliminating trajectory bias), while Cousin data uses scripted skills to autonomously enrich state-action diversity.

Deployment and Computational Cost. Deployment requires a panoramic camera for real scene capture. In simulation, collecting 50 automated trajectories within the digital cousins takes approximately 30 minutes on a single NVIDIA RTX 4090 GPU, with VRAM usage peaking at 6GB.

B. Comparison with Related Simulation Frameworks

Table VI systematically evaluates our method against state-of-the-art baselines across six critical dimensions: **Real2Sim2Real** capability, **Visual Fidelity** (e.g., 3DGS photorealism), **Physics Fidelity Support** (spanning rigid, articulated, deformable, and fluid dynamics), **Digital Cousins** generation (systematic scene variations), **Auto Scene Generation** (manual-free scene construction), and **Auto Data Collection** (teleop-free trajectory generation).

Unlike prior works that often compromise on visual realism, physical complexity, or scalability, our framework uniquely synergizes all six capabilities, providing a comprehensive and automated pipeline for robust robot learning.

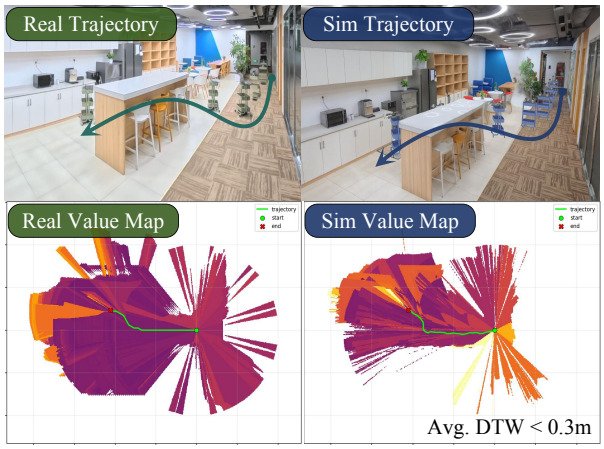

Fig. 9: Real world vs. Simulation navigation trajectories and value maps.

TABLE V: Zero-shot navigation success rate in multi-room environments.

Scene	Success Rate (SR)
Sim	12/20
Real	10/20

C. Additional Task Visualization

Due to space limitations in the main manuscript, we only visualized a subset of the simulation tasks. In this section, we present the visualizations for the remaining tasks supported by our pipeline: *Set Tableware*, *Close Drawer*, and *Assemble Burger*.

As shown in Fig. 10, for each task, we display the **Digital Twin** (a precise reconstruction of the real-world setup) alongside a generated **Digital Cousin** (a semantically equivalent but visually diverse variation). These qualitative results further demonstrate the capabilities of our pipeline in generating high-fidelity and diverse simulation environments for a wide range of manipulation tasks.

D. Detailed Quantitative Evaluation

In this section, we provide the detailed quantitative results corresponding to the simulation-to-real correlation analysis presented in the main text. We report the detailed success rates for all baselines (ACT, DP, SmolVLA and π_0) across three manipulation tasks: *Set Tableware*, *Open Microwave*, and *Fold Cloth*.

Experimental Settings. To ensure a rigorous and fair comparison, we standardized the data collection and evaluation protocols for both simulation and real-world experiments:

TABLE VI: Structural comparison against related simulation frameworks.

Method	Real2Sim2Real	Visual Fidelity	Physics Fidelity Support	Digital Cousins	Auto Scene Gen.	Auto Data Collect
RoboGen [51]	×	×	Rigid & Articulated	×	✓	✓
SaR Co-Training [34]	×	×	Rigid & Articulated	✓	×	×
URDFormer [8]	✓	×	Rigid & Articulated	×	✓	×
GSWorld [22]	✓	✓	Rigid	×	✓	×
SplatSim [42]	✓	✓	Rigid	×	✓	✓
Ours (WorldComposer)	✓	✓	Rigid, Articulated, Deformable, Fluid	✓	✓	✓

Fig. 10: **Additional Simulation Tasks.** Visualization of the *Digital Twin* (left) and generated *Digital Cousin* (right) environments for the tasks not displayed in the main manuscrip.

- **Simulation:** We utilized a dataset comprising **100** expert trajectories per task for training. During the evaluation phase, each policy was tested over **100** independent trials for each generalization setting (Train, Unseen Scene, Unseen Object, and Unseen Scene & Object) to report the average success rate.
- **Real-World:** We collected **50** high-quality human teleoperated trajectories for each task. For the real-robot evaluation, considering the operational costs, we conducted **20** evaluation trials for each method under each specific setting.

The specific breakdown of performance metrics is organized as follows: Table VII, VIII, IX present the simulation results, while Table X, XI, XII detail the corresponding real-world performance.

TABLE VII: Simulation Evaluation (Set Tableware).

Method \ Task Level	Train	Scene	Object	S & O
ACT	0.83	0.54	0.36	0.13
DP	0.69	0.61	0.30	0.25
SmolVLA	0.85	0.78	0.49	0.40
π_0	0.91	0.84	0.53	0.51

TABLE VIII: Simulation Evaluation (Open Microwave).

Method \ Task Level	Train	Scene	Object	S & O
ACT	0.91	0.61	0.46	0.23
DP	0.64	0.59	0.33	0.25
SmolVLA	0.82	0.76	0.49	0.44
π_0	0.89	0.82	0.58	0.54

TABLE IX: Simulation Evaluation (Fold Cloth).

Method \ Task Level	Train	Scene	Object	S & O
ACT	0.58	0.33	0.35	0.19
DP	0.53	0.45	0.29	0.22
SmolVLA	0.65	0.61	0.44	0.38
π_0	0.71	0.67	0.55	0.50

TABLE X: Real-World Evaluation (Set Tableware).

Method \ Task Level	Train	Scene	Object	S & O
ACT	0.70	0.25	0.40	0.05
DP	0.60	0.30	0.30	0.10
SmolVLA	0.95	0.85	0.70	0.55
π_0	0.95	0.90	0.90	0.85

TABLE XI: Real-World Evaluation (Open Microwave).

Method \ Task Level	Train	Scene	Object	S & O
ACT	0.90	0.45	0.60	0.25
DP	0.75	0.55	0.50	0.35
SmolVLA	1.00	0.95	0.80	0.75
π_0	1.00	0.95	0.85	0.85

TABLE XII: Real-World Evaluation (Fold Cloth).

Method \ Task Level	Train	Scene	Object	S & O
ACT	0.50	0.15	0.10	0.00
DP	0.45	0.20	0.05	0.00
SmolVLA	0.45	0.35	0.15	0.05
π_0	0.65	0.60	0.35	0.35

E. More Experiments

In this section, we provide a more comprehensive evaluation of our framework by presenting additional Domain Randomization comparisons, detailed ablation studies, and scaling analyses across a broader range of tasks.

Effectiveness over Domain Randomization. As shown in Table XIII, while standard Domain Randomization (Lighting, Texture) yields moderate gains, it cannot alter scene structures.

TABLE XIII: Additional baselines (Domain Randomization) & Ablations.

	Set Tableware		Open Microwave		Fold Cloth		Average	
	Scene	Object	Scene	Object	Scene	Object	Scene	Object
50 Real + 50 Sim DR (Lighting)	0.40	0.40	0.60	0.55	0.40	0.15	0.47	0.37
50 Real + 50 Sim DR (Texture)	0.30	0.45	0.65	0.45	0.30	0.20	0.42	0.37
50 Real + 50 Sim Cousin (Scene)	0.50	0.40	0.70	0.50	0.40	0.25	0.53	0.38
50 Real + 50 Sim Cousin (Object)	0.30	0.45	0.55	0.60	0.35	0.30	0.40	0.45
50 Real + 50 Sim Cousin (WorldComposer)	0.55	0.50	0.70	0.60	0.40	0.35	0.55	0.48

TABLE XIV: Scaling curves on more tasks (0^*: 50 real data only).

Added Cousin Sim Data	0^*	50	100	200	500	1000
Open Microwave	0.20	0.45	0.60	0.75	0.85	0.85
Fold Cloth	0.05	0.30	0.40	0.50	0.55	0.55

Ablating variation types shows both Cousin (Scene & Object) independently outperform classical DR. Integrating them via WorldComposer achieves peak performance, clarifying that our gains stem from scene and object diversity, not sheer data volume or simple texture randomization.

Scaling Efficiency on Additional Tasks. We provide per-task scaling curves for two additional tasks (Table XIV). Both demonstrate strong, consistent data efficiency, with performance reliably scaling up as simulated cousins increase.