GROUNDED-DREAMER: ROBOT LEARNING FROM WORLD MODEL SYNTHETIC DATA WITH PHYSICAL GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Scaling robot manipulation policies requires large amounts of diverse, action-labeled data, yet collecting such data in the real world remains expensive and difficult to scale. Existing methods either depend on human teleoperation, rely on simulation with limited visual realism, or use human videos that lack embodiment-specific annotations. In this paper, we introduce Grounded-Dreamer, a framework for synthesizing physically grounded robot manipulation data with video world models while avoiding human teleoperation demonstrations. Grounded-Dreamer combines the visual realism of generative video models with simulator physics engines. Starting from a small set of real images depicting robot task initial states, our method automatically reconstructs task-relevant objects and camera geometry, aligns them in simulation, and collects physically valid trajectories. These trajectories are used to post-train both a video world model and an inverse dynamics model, enabling the generation of visually realistic robot manipulation videos and corresponding robot actions from real images. To filter physically implausible rollouts, Grounded-Dreamer replays generated trajectories in the reconstructed simulation and retains only those that successfully complete the tasks. The real-world robot policies are trained solely on the grounded synthetic trajectories without any human demonstrations. Empirical results in real-world tasks demonstrate that Grounded-Dreamer can synthesize physically grounded data and substantially facilitate policy learning without requiring human teleoperation effort. Project Page: `https://sites.google.com/view/groundeddreamer-project`.

1 INTRODUCTION

Building general-purpose robot policies capable of performing diverse manipulation tasks across varied environments requires large-scale, high-quality data. However, robot data is difficult to scale, and existing sources present a challenging trade-off between quality and quantity. Real-world teleoperation provides high-fidelity demonstrations but is limited by intensive human labor. While UMI data collection (Chi et al., 2024; Xu et al., 2025) improves efficiency, there remains a need for manual operations. Conversely, simulation data offers inexpensive, action-labeled trajectories but is hindered by the sim-to-real visual gap. Human-centric videos are abundant and diverse, yet they lack the robot action annotations essential for direct policy training. Consequently, the field remains trapped between costly manual collection and the limitations of imperfect proxy data.

This dilemma motivates the pursuit of scalable robot data synthesis. Prior efforts such as DemoGen (Xue et al., 2025) and MimicGen (Mandlekar et al., 2023) have attempted to augment existing datasets by diversifying robot demonstrations; however, these methods still rely on human-collected trajectories as initial seeds. Other approaches, including RoboGen (Wang et al., 2024b) and GenSim (Wang et al., 2024a; Hua et al., 2025), leverage Large Language Models (LLMs) and simulators to automate task generation and trajectory collection. While these frameworks significantly reduce human intervention, the resulting data remains confined to simulation, leaving the visual sim-to-real discrepancy unresolved.

Generative video world models offer a potential alternative. Trained on internet-scale data, these models capture rich visual and world dynamics priors, enabling the generation of manipulation

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

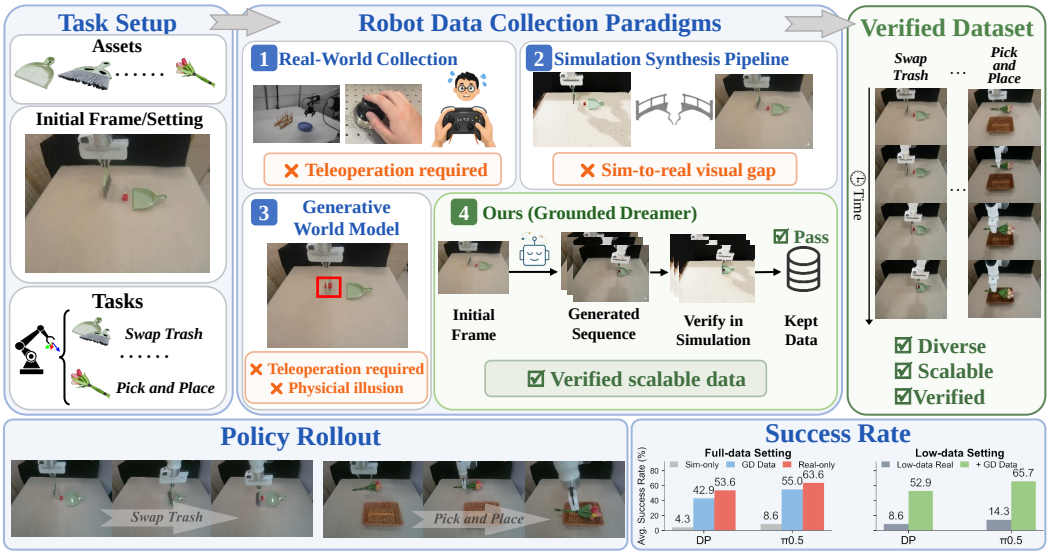

Figure 1: **Overview of Grounded-Dreamer.** Our framework automates the generation of massive, high-fidelity robot data from a small number of real-world reference images. By combining generative world models with simulation-based verification, Grounded-Dreamer produces data that is both visually realistic and physically consistent. This simulation verification loop effectively filters out the physical hallucinations common in purely generative approaches, allowing robot policies to be trained directly on scalable synthetic data that transfers robustly to complex real-world scenarios.

rollouts. When paired with an Inverse Dynamics Model (IDM), such videos can be converted into action-labeled trajectories. However, pretrained models like Cosmos (NVIDIA et al., 2025) are primarily trained on human-centric videos. Directly applying them to robotic tasks often results in "hallucinated" physics, such as inconsistent object motions or non-physical contacts.

In this work, we introduce **Grounded-Dreamer**, a framework that synthesizes high-fidelity robot manipulation data without a single human teleoperation demonstration. Our core insight is to link the visual realism of generative world models with the grounding of physics simulators. We first build an automatic real-to-sim pipeline that calibrates camera poses, reconstructs task-relevant objects, and aligns them with the scene. Within this reconstructed simulator, we automatically collect diverse trajectories by leveraging the simulator's ground-truth states. These grounded trajectories serve as the foundation for post-training both the video world model and the IDM.

During data synthesis, the post-trained world model generates visually realistic manipulation videos from real-world initial images, while the IDM extracts the corresponding robot actions. To guarantee physical validity, we implement a closed-loop verification step: replaying the inferred trajectories within the reconstructed simulation. This process filters out physically inconsistent generations, ensuring that only successful, grounded trajectories are used for further policy training.

We evaluate Grounded-Dreamer across seven diverse real-world manipulation tasks, spanning short-to-long horizon sequences and sophisticated tool use. Our empirical results demonstrate that policies trained exclusively on grounded synthetic data achieve performance competitive with those trained on real-world human demonstrations. Furthermore, owing to the inherent visual diversity and our targeted augmentation strategies, Grounded-Dreamer significantly outperforms real-data policies in out-of-distribution scenarios, showcasing robust zero-shot generalization across novel scenarios.

Our contributions are summarized as follows:

- We propose **Grounded-Dreamer**, a data synthesis framework that eliminates the need for human demonstrations by linking video world models with simulator physics engines.
- We introduce an automated real-to-sim reconstruction and collection pipeline, enabling data synthesis that maintains both high visual fidelity and strict physical grounding.

- We demonstrate through extensive real-world experiments that our grounded synthetic data effectively facilitates policy learning, achieving robust performance across diverse scenarios through purely autonomous data generation.

2 Related Work

2.1 Robot Data Synthesis

Simulation has been widely used to generate diverse conditions for robot data. Early sim-to-real work relied on visual or dynamics randomization (Tobin et al., 2017; Peng et al., 2018). Later environments extended this idea to broader robot-learning settings: RLBench standardizes manipulation tasks (James et al., 2020), robosuite and SAPIEN provide modular robot-object simulation platforms (Zhu et al., 2020; Xiang et al., 2020), and Isaac Gym and RoboCasa emphasize scalable physics or household manipulation (Makoviychuk et al., 2021; Nasiriany et al., 2024).

Recent generative frameworks, such as GenSim, GenSim2, and RoboGen, use language or foundation models to generate tasks, scenes, demonstrations, or supervision (Wang et al., 2024a; Hua et al., 2025; Wang et al., 2024b). Their data quality, however, still depends on generated task programs, planners, solvers, or simulator assumptions. For manipulation, local grounding outweighs global visual realism. Incorrect object pose, axis, or camera frame can make a plausible trajectory physically misleading. 6D pose estimation and reconstruction recover task-relevant structure from real observations (Hodan et al., 2018; Wen et al., 2024), while structure-from-motion, NeRF, and 3D Gaussian Splatting facilitate view-consistent geometry or appearance (Schönberger & Frahm, 2016; Mildenhall et al., 2020; Kerbl et al., 2023).

Another route reuses collected trajectories. A series of studies adapt human demonstrations to new object configurations (Mandlekar et al., 2023), corrective interventions (Hoque et al., 2024), long-horizon skills (Garrett et al., 2025), and bimanual dexterous manipulation (Jiang et al., 2025). DemoGen further adapts one demonstration and edits the 3D scene to synthesize new observations (Xue et al., 2025). DROID and Open X-Embodiment instead scale real robot data directly, but require substantial collection effort (Khazatsky et al., 2024; O'Neill et al., 2024). Grounded-Dreamer lies between these directions: it grounds synthetic data in real scene-object inputs, preserves physical structure through simulation, and expands the distribution with world-model rollouts.

2.2 World Models for Robot Learning

World models were originally developed as learned dynamics models for predicting future states before acting (Ha & Schmidhuber, 2018; Hafner et al., 2020; 2025). In robotics, visual foresight leveraged video prediction for model-based control by selecting actions whose predicted futures match a goal (Finn & Levine, 2017). More recently, diffusion-based video models have enabled realistic rollout generation from images or text (Ho et al., 2022), while systems such as UniSim and Genie move toward controllable visual worlds (Yang et al., 2024; Bruce et al., 2024).

However, visually plausible videos are not directly usable robot training trajectories, as they often lack actions, embodiment constraints, and physically reliable interactions. DreamGen adapts video world models to robot learning by recovering pseudo actions through latent action or inverse dynamics models (Jang et al., 2025), while GigaBrain-0 further scales world-model-generated data for VLA training (GigaBrain Team, 2025). Prior works have also explored inverse dynamics models to bridge observations and actions (Agrawal et al., 2016; Pathak et al., 2018). Once converted into action-conditioned trajectories, such generated data can train policies including Diffusion Policy, RT-2, and π_0 (Chi et al., 2025; Zitkovich et al., 2023; Black et al., 2024). Grounded-Dreamer builds on this line of work by coupling world-model rollouts with reconstructed scene inputs and simulation-based grounding to ensure both visual realism and physical consistency before policy learning.

3 Methodology

We introduce **Grounded-Dreamer**, an autonomous pipeline designed to generate verifiable, high-fidelity synthetic data for scalable robot policy learning without the need for human-collected real-world trajectories. As illustrated in Figure 2, starting from a small set of reference images captured

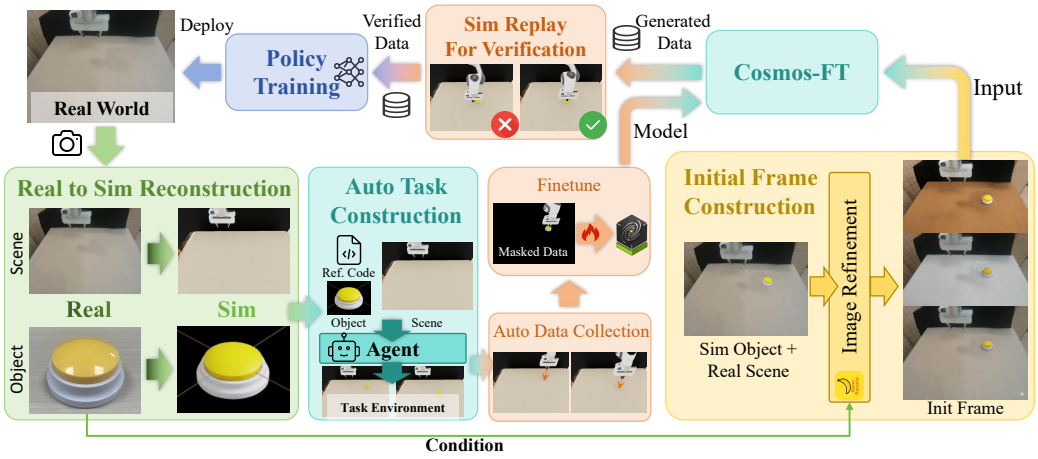

Figure 2: **The Grounded-Dreamer Pipeline.** Given several real-world reference images, our pipeline first reconstructs physically aligned simulation scenes and objects. Within these environments, task data is automatically collected and processed with interaction-aware masking to fine-tune Cosmos-Predict 2.5 as the core data generation model. An image refinement model then synthesizes diverse initial frames by combining reconstructed assets with real scene backgrounds and object appearances. Conditioned on these frames, the generation model produces manipulation videos, which are converted into action trajectories using an Inverse Dynamics Model (IDM). Finally, all synthesized trajectories are replayed in simulation for closed-loop verification, filtering out physically inconsistent samples before policy training and real-world deployment.

in the real world, Grounded-Dreamer synthesizes massive, validated datasets through an automated workflow, which consists of four core stages: (i) **Task-Oriented Simulation Reconstruction**, where we reconstruct physically consistent simulation scenes and task-relevant objects; (ii) **Sim Data Collection and Interaction-Aware Masking**, where simulation data is automatically collected to fine-tune Cosmos-Predict 2.5 and the IDM with interaction-aware masking; (iii) **Initial Frame Synthesis**, where the Nano Banana-based pipeline aligns real-sim composites and generates augmented initial images to enhance diversity and visual fidelity; (iv) **Data Synthesis and Closed-Loop Verification**, where synthetic trajectories are generated from real-world initial frames and validated via simulation verification.

3.1 Task-Oriented Simulation Reconstruction

Scene Reconstruction. Our framework is built upon the Lehome simulation environment (Li et al., 2026). Given the workspace boundaries and the relative pose of the robot base, we construct a lightweight simulation scene containing the robotic arm and workspace table. To align the simulated and real-world viewpoints, we perform extrinsic calibration using the FEEPE framework (Wu et al., 2025) and directly adopt the intrinsic parameters of the physical camera. This process ensures consistent fields of view (FOVs) between simulation and reality, providing a reliable geometric foundation for data grounding. For efficiency, we reuse off-the-shelf Lehome assets for tabletop textures and environmental backgrounds, since pixel-level fidelity in task-irrelevant regions is unnecessary for our pipeline.

Task-Oriented Object Reconstruction. For **rigid objects**, we employ SAM3D to reconstruct 3D geometry from the initial observation frame and rescale the resulting meshes to match real-world dimensions (SAM 3D Team et al., 2025). For **articulated objects**, we first use SAM3D to decompose the object into individual components (SAM 3D Team et al., 2025). Our autonomous agent then takes these components, together with predefined connectivity rules and joint specifications, and automatically assembles the articulated kinematic structure within the simulator.

3.2 SIM DATA COLLECTION AND INTERACTION-AWARE MASKING

With the task-aligned simulation environment established, we develop an automated data collection pipeline for trajectory generation necessary for world model post-training. Instead of manually scripting each task, we employ an LLM-based agent to synthesize and execute data collection scripts from three inputs: (i) **Reference Code Templates** defining trajectory generation logic, including modular execution stages and inverse kinematics (IK)-based end-effector control; (ii) **High-level Task Descriptions** specifying task objectives and object semantics; and (iii) **Object Metadata** automatically retrieved from Isaac Sim (NVIDIA, 2023), including object poses and physical properties. Using this framework, we collect 100-150 trajectories for each task, with the volume adjusted according to task complexity and object count.

To reduce the sim-to-real domain gap, we propose an **Interaction-Aware Masking** strategy that preserves only the robotic arm and interacting objects in simulation frames (Figure 3 and 8). This preprocessing encourages Cosmos-Predict 2.5 (NVIDIA et al., 2025) and IDM (Jang et al., 2025; Baker et al., 2022) to focus on modeling interaction dynamics during fine-tuning. As a result, when conditioned on real-world initial frames, the fine-tuned Cosmos model generates realistic robot-object interactions while maintaining stable task-irrelevant backgrounds through its generative priors.

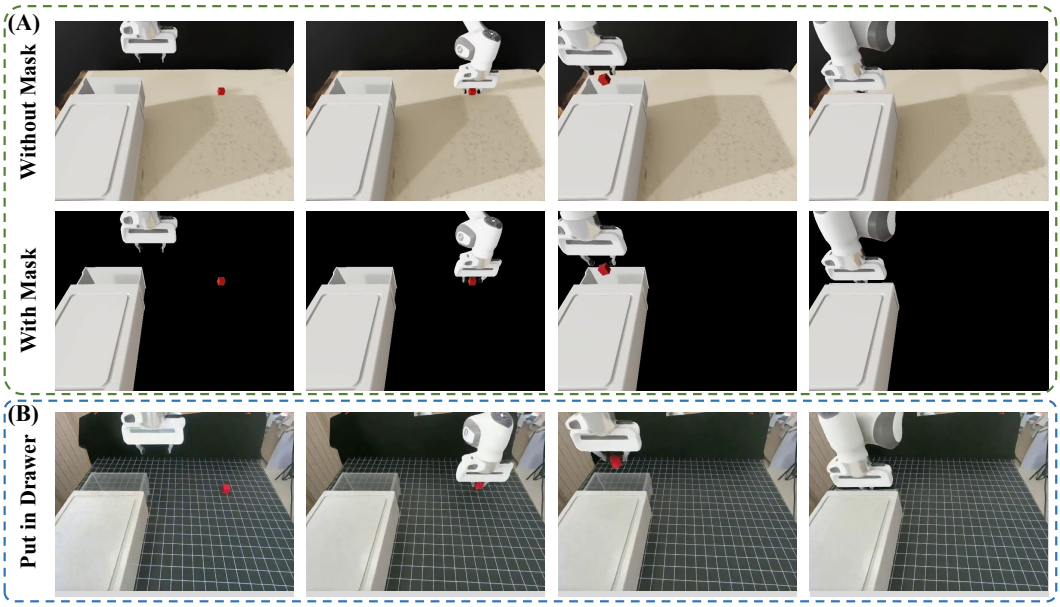

Figure 3: Qualitative results of (A) simulation data acquisition and interaction-aware masking, and (B) Grounded-Dreamer video generation.

3.3 INITIAL FRAME SYNTHESIS

We can leverage the fine-tuned video world model to generate new rollouts conditioned on task-specific initial frames. To increase data diversity, we randomize object poses in the simulated workspace and render corresponding RGB images, instance masks, and object coordinates. We then construct preliminary scenes by compositing masked simulated objects onto real-world background images. To bridge the visual domain gap, Nano Banana 2 leverages reference photographs of the real objects to transform the composites into photorealistic images (Google DeepMind, 2026). We further exploit its generative capability to augment the initial frames with variations in lighting conditions and tablecloth textures, improving robustness to distribution shifts. Additional implementation details and visualizations are provided in Appendix D.

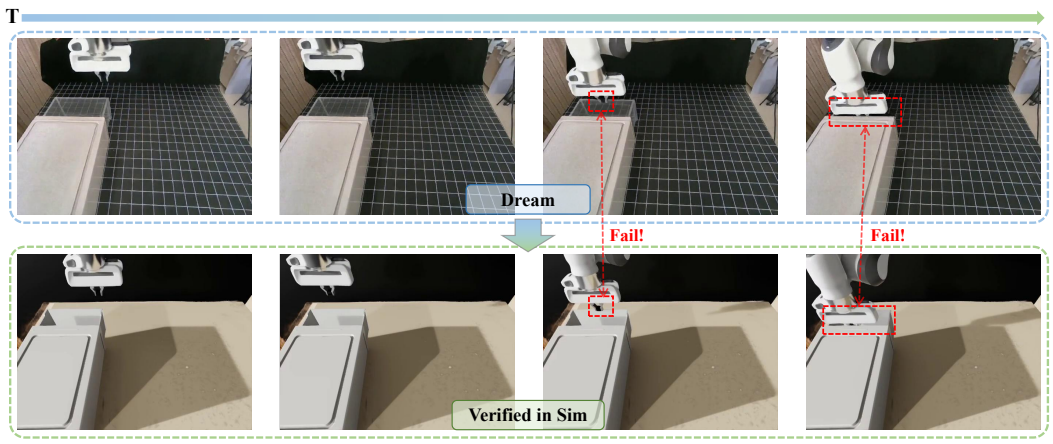

Figure 4: Visualization of Simulation Validation. The physical simulation correctly captures a task failure due to incorrect gripper placement, whereas the world model exhibits spatial hallucination by generating a hallucinated successful outcome.

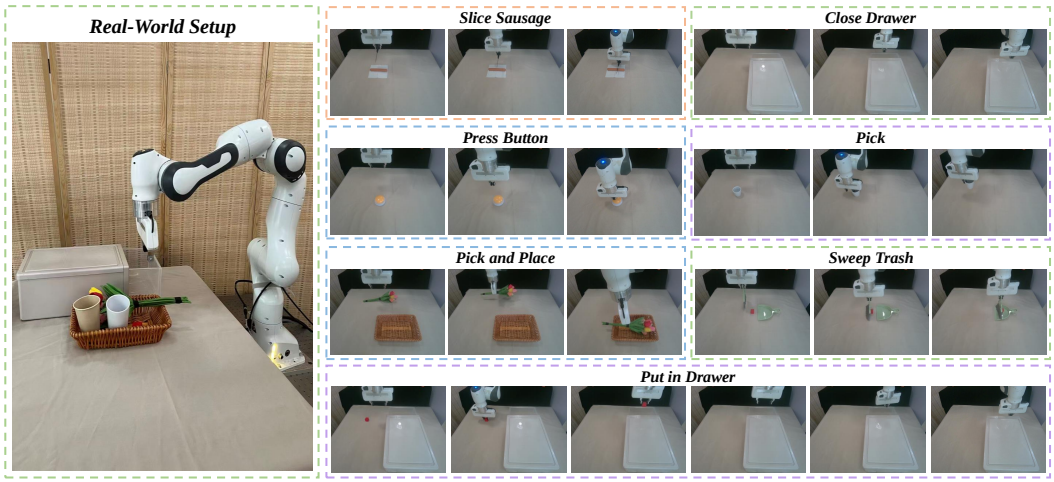

Figure 5: **Real-world Setup and Task Execution.** (Left) Real-robot evaluation environment; (Right) Visualizations of our policy executing manipulation tasks.

3.4 Data Synthesis and Closed-Loop Verification

Conditioned on the augmented initial frames, the fine-tuned Cosmos-Predict 2.5 model generates rollout videos (Figure 3 & 9. We apply Interaction-Aware Masking using SAM (Kirillov et al., 2023) and employ the IDM to extract action trajectories, forming paired synthetic training data.

However, generative world models may produce **physical hallucinations**, including (1) **Morphological Hallucinations**, such as invalid robot kinematics or physically infeasible joint configurations; (2) **Spatial-Depth Ambiguities**, where apparent 2D contacts do not reflect valid 3D interactions.

To address these issues, we introduce a **Closed-Loop Verification** mechanism. Using the reconstructed simulation environment and initialized object poses (Sec. 3.3), we replay synthesized action trajectories in the physics simulator and classify them as *Optimal* or *Sub-optimal* according to task success criteria. Figure 4 illustrates this process. This verification stage filters out physically inconsistent or task-failing trajectories, ensuring high-quality data for downstream policy learning. Experiments in Sec. 4 show that this filtering is critical for maintaining strong policy performance.

Table 1: **Main experimental results across seven tasks.** We compare the success rates of DP and $\pi_{0.5}$ under different data regimes, with RIGVid as a representative baseline.

Method	Pick	Button	Close Drawer	Pick and Place	Sweep Trash	Slice Sausage	Put in Drawer
Sim-only (DP)	10%	10%	0%	5%	5%	0%	0%
Real-only (DP)	50%	60%	**60%**	**50%**	50%	**70%**	35%
GD data (DP)	*45%*	*50%*	*45%*	*35%*	*45%*	*55%*	*25%*
Low-data Real (DP)	10%	10%	15%	0%	10%	10%	5%
Co-training (DP)	**55%**	**65%**	50%	45%	**55%**	60%	**40%**
Sim-only ($\pi_{0.5}$)	10%	15%	*15%*	5%	10%	5%	0%
Real-only ($\pi_{0.5}$)	65%	75%	**65%**	**60%**	60%	**75%**	**45%**
GD data ($\pi_{0.5}$)	*55%*	75%	55%	*50%*	60%	*60%*	30%
Low-data Real ($\pi_{0.5}$)	20%	20%	10%	15%	20%	10%	5%
Co-training ($\pi_{0.5}$)	**70%**	**80%**	**65%**	55%	**75%**	70%	**45%**
RIGVid	25%	20%	5%	5%	0%	0%	0%

Table 2: **Success Rates in OOD Scenarios.** Performance comparison of models trained on two types of data across seven manipulation tasks under out-of-distribution (OOD) conditions.

Method	Pick	Button	Close Drawer	Pick and Place	Sweep Trash	Slice Sausage	Put in Drawer
Real-only (DP)	10%	5%	5%	0%	5%	5%	0%
Verified WM data (DP)	**40%**	**50%**	**35%**	**30%**	**40%**	**45%**	**15%**
Real-only ($\pi_{0.5}$)	25%	20%	10%	20%	20%	15%	5%
Verified WM data ($\pi_{0.5}$)	**45%**	**60%**	**55%**	**40%**	**60%**	**50%**	**30%**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Manipulation Tasks. We evaluate Grounded-Dreamer across seven diverse real-world tasks: *Pick*, *Close Drawer*, *Press Button*, *Pick and Place*, *Put in Drawer*, *Sweep Trash*, and *Slice Sausage*. This suite covers a broad spectrum of requirements, ranging from short-horizon object interactions to sophisticated tool usage and handling of deformable objects. Crucially, our evaluation incorporates both in-distribution benchmarks and challenging Out-of-Distribution scenarios, providing a rigorous assessment of the framework's zero-shot generalization across novel environments and unseen object configurations.

Hardware Setup. The physical experimental platform centers on a Franka Emika robot arm paired with a Panda gripper. Visual observations are captured by an Intel RealSense D405 depth camera. For real-time inference and control, the trained policies are deployed on a workstation equipped with an NVIDIA GeForce RTX 4090 GPU. Figure 5 illustrates the comprehensive hardware configuration alongside representative real-world execution trajectories.

Baselines and Metrics. We employ Diffusion Policy (DP) (Chi et al., 2025) and $\pi_{0.5}$ (Physical Intelligence et al., 2025) as representative policy backbones. To provide a comprehensive evaluation of Grounded-Dreamer, we compare our approach against several data regimes and alternative methodologies: (i) **Sim-only**: Policies trained exclusively on raw simulation data, representing a performance lower bound; (ii) **Real-only**: Policies trained on a full set of 50 human-labeled real-world trajectories per task; (iii) **Low-data Real**: A regime where only 5 real-world samples are available, representing extreme data scarcity; (iv) **RIGVid** (Patel et al., 2025): A baseline that utilizes a video world model to predict object-centric hand-object interactions, from which end-effector poses are extracted and mapped to robot actions.

The primary evaluation metric is the **Success Rate**. We perform 30 trials per task for scaling law analysis and 20 trials for all other comparative evaluations.

4.2 Main Results and Comparative Analysis

We evaluate the performance of DP and $\pi_{0.5}$ models trained with different data regimes (50 trajectories per task). As summarized in Table 1, the **Sim-only** baseline performs poorly on real-world tasks, failing to complete most interactions due to significant visual and physical domain gaps. In contrast, the *Real-only* baseline maintains a high success rate. Notably, policies trained exclusively on our **verified GD data** achieve competitive performance relative to the *Real-only* baseline. Specifically, for the *Pressing Button* task, the $\pi_{0.5}$ policy reaches parity with the performance of its real-world counterpart. This demonstrates the immense potential of world model data, particularly for labor-intensive tasks like *Sweeping Trash* and *Slicing Sausage*, where synthetic data can be generated autonomously from few frames at near-zero cost. We also compare our framework with RIGVid, which utilizes world models through an alternative paradigm. RIGVid performs significantly below expectations, exhibiting a substantial performance gap compared to both the Real-only and our verified GD-data regimes. This performance degradation is primarily attributed to inaccuracies in hand-pose extraction and susceptibility to viewpoint jitter.

Furthermore, we investigate Co-training using a minimal set of real demonstrations combined with 50 GD trajectories. While *Low-data Real* models fail to generalize across backbones—yielding only 8.6% for DP and 14.3% for $\pi_{0.5}$—adding GD data leads to a substantial performance leap. This regime not only outperforms the *GD-data-only* model but also matches or exceeds the 50-sample *Real-only* baseline. Here, the few real trajectories act as **domain anchors**, aligning the synthetic distribution with the real domain and further unlocking the world model's latent potential.

4.3 Generalization and Robustness to Environment Variation

To enhance environmental diversity, we utilize the *Nano Banana 2* model (Google DeepMind, 2026) to augment real-world initial frames with varying backgrounds, tablecloth textures, and lighting conditions before generating 50 GD trajectories. We then evaluate the DP and $\pi_{0.5}$ models in **Out-of-Distribution (OOD)** scenarios. As shown in Table 2, models trained with GD data exhibit superior generalization capabilities, maintaining robust performance in novel environments. Conversely, the *Real-only* baseline, constrained by the limited diversity of its training set, fails to generalize; its average success rates drop to only **7.1% and 16.4% for DP and $\pi_{0.5}$, respectively.** Crucially, this environmental variety is achieved autonomously through image-to-image augmentation of initial frames, requiring no additional physical intervention. Attaining a comparable level of diversity through manual real-world collection would necessitate exhaustive and repetitive re-collection cycles, incurring prohibitive overhead in terms of human labor and physical resources.

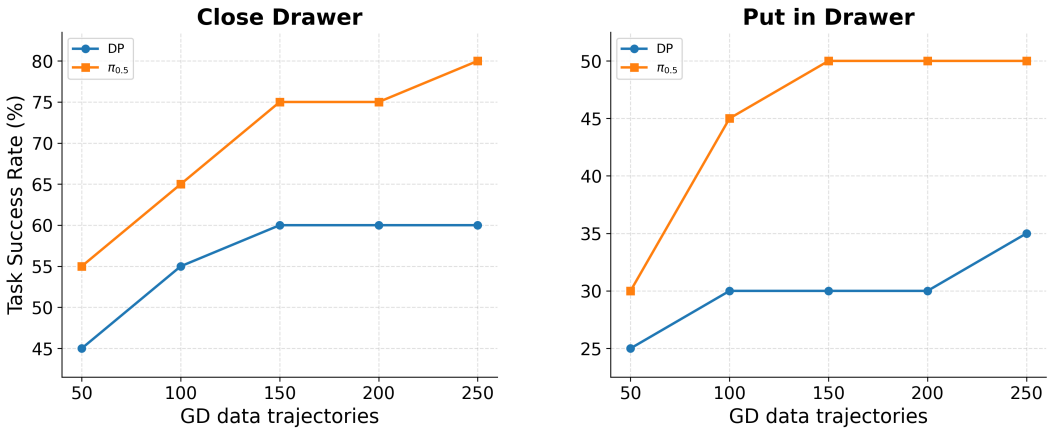

Figure 6: **Scaling Success Rates.** Success rates on two representative tasks improve as the number of Grounded-Dreamer (GD) generated trajectories increases.

4.4 Scaling Laws of Synthetic World Model Data

To evaluate the efficiency and potential ceiling of training with synthetic data, we investigate the scaling laws of GD-generated trajectories in the *Close Drawer* and *Put in Drawer* tasks. As illustrated in Figure 6, increasing the number of synthetic trajectories improves success rates for both tasks. However, performance begins to plateau at approximately 150 trajectories, with further increases yielding only marginal improvements.

Table 3: Ablation on Interaction-Aware Masking.

Method	Pick	Button	Close Drawer	Pick and Place	Sweep Trash	Slice Sausage	Put in Drawer
w/ Masking (DP)	**45%**	**50%**	**45%**	**35%**	**45%**	**55%**	**25%**
w/o Masking (DP)	0%	5%	0%	0%	0%	0%	0%
w/ Masking ($\pi_{0.5}$)	**55%**	**75%**	**55%**	**50%**	**60%**	**60%**	**30%**
w/o Masking ($\pi_{0.5}$)	5%	0%	10%	0%	10%	5%	0%

Table 4: Ablation on Replay Verification.

Method	Pick	Button	Close Drawer	Pick and Place	Sweep Trash	Slice Sausage	Put in Drawer
100% (DP)	**45%**	**50%**	**45%**	**35%**	**45%**	**55%**	**25%**
80% (DP)	40%	40%	35%	30%	30%	45%	10%
50% (DP)	25%	30%	25%	25%	20%	20%	5%
100% ($\pi_{0.5}$)	**55%**	**75%**	**55%**	**50%**	**60%**	**60%**	**30%**
80% ($\pi_{0.5}$)	35%	60%	50%	35%	45%	50%	20%
50% ($\pi_{0.5}$)	25%	35%	25%	20%	25%	30%	10%

4.5 Ablation Studies: Key Components of Grounded-Dreamer

In this section, we evaluate the individual contributions of two critical components of our framework: **Interaction-Aware Masking** and **Closed-Loop Replay Verification**.

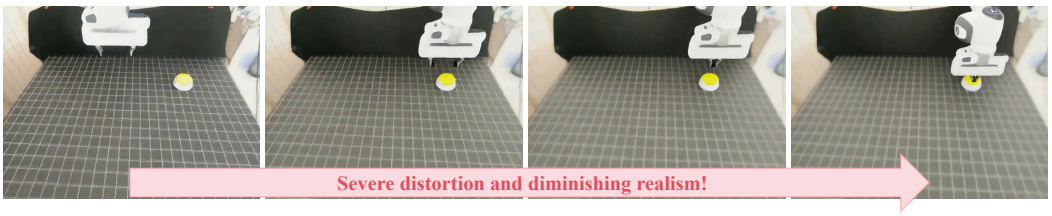

Figure 7: **Ablation without Masking.** Visualizations of video rollouts generated by the world model fine-tuned on raw simulation data without the interaction-aware masking strategy.

Effect of Interaction-Aware Masking. We compare downstream policies trained on synthetic data from world models fine-tuned with either interaction-aware masks or raw simulation data. Without masking, generated videos exhibit distorted pixels and unrealistic scene deformations (Fig. 7), and success rates drop significantly (Table 3). These results support masking task-irrelevant regions during world-model fine-tuning.

Necessity of Closed-Loop Replay Verification. Table 4 compares policies trained with varying ratios of verified (*Optimal*) versus non-verified (*Sub-optimal*) trajectories. Task success increases monotonically with the proportion of verified data. Notably, at 50% sub-optimal data, performance **suffers a precipitous drop**. These results confirm the importance of simulation replay for filtering synthetic training data.

AI USE STATEMENT

We use SAM3D for object reconstruction and an LLM-based agent for simulation task construction and data-collection script generation. We use Nano Banana 2 to synthesize and augment initial frames, and Cosmos-Predict 2.5 with an inverse dynamics model to generate synthetic robot training data. Generated action trajectories are replayed in simulation and filtered according to task success, as described in Section 3. The authors take responsibility for the final manuscript and AI-assisted artifacts.

ETHICS STATEMENT

This work studies robot manipulation and does not involve human or animal subjects. Synthetic data may inherit biases and physical inaccuracies from generative models. Our simulation verification filters generated trajectories but does not guarantee safe deployment in all real-world settings. We therefore recommend task-specific safety evaluation and appropriate safeguards before deployment, particularly in human environments. Broader impacts and limitations are discussed in the appendix.

REPRODUCIBILITY STATEMENT

To support reproducibility, Section 3 describes the reconstruction, data generation, and verification pipeline, and Section 4 specifies the experimental setup, baselines, and evaluation protocol. Appendix D provides the inputs and prompting procedure for initial-frame synthesis, while Appendix E reports model training and inference settings. Additional qualitative examples are provided in Appendix C.

REFERENCES

Pulkit Agrawal, Ashvin Nair, Pieter Abbeel, Jitendra Malik, and Sergey Levine. Learning to poke by poking: Experiential learning of intuitive physics. In *Advances in Neural Information Processing Systems*, 2016.

Bowen Baker, Ilge Akkaya, Peter Zhokhov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video pretraining (VPT): Learning to act by watching unlabeled online videos. *Advances in Neural Information Processing Systems*, 35:24639–24654, 2022.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A Vision-Language-Action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Ria Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie C. Y. Chan, Nicolas Heess, Luis Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, and Tim Rocktäschel. Genie: Generative interactive environments. In *International Conference on Machine Learning*, 2024.

Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. *arXiv preprint arXiv:2402.10329*, 2024.

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*, 44(10–11):1684–1704, 2025. doi: 10.1177/02783649241273668.

Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *IEEE International Conference on Robotics and Automation*, pp. 2786–2793, 2017.

Caelan Garrett, Ajay Mandlekar, Bowen Wen, and Dieter Fox. SkillMimicGen: Automated demonstration generation for efficient skill learning and deployment. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2750–2790, 2025.

GigaBrain Team. GigaBrain-0: A world model-powered vision-language-action model. *arXiv preprint arXiv:2510.19430*, 2025.

Google DeepMind. Gemini 3.1 Flash Image (Nano Banana 2): Model card. Technical report, Google DeepMind, 2026. URL `https://deepmind.google/models/model-cards/gemini-3-1-flash-image/`. Updated 26 February 2026; Accessed: 2026-05-07.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025. doi: 10.1038/s41586-025-08744-2.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Video diffusion models. In *Advances in Neural Information Processing Systems*, 2022.

Tomas Hodan, Frank Michel, Eric Brachmann, Wadim Kehl, Anders Glent Buch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, Caner Sahin, Fabian Manhardt, Federico Tombari, Tae-Kyun Kim, Jiri Matas, and Carsten Rother. BOP: Benchmark for 6D object pose estimation. In *European Conference on Computer Vision Workshops*, 2018.

Ryan Hoque, Ajay Mandlekar, Caelan Garrett, Ken Goldberg, and Dieter Fox. IntervenGen: Interventional data generation for robust and data-efficient robot imitation learning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2024.

Pu Hua, Minghuan Liu, Annabella Macaluso, Yunfeng Lin, Weinan Zhang, Huazhe Xu, and Lirui Wang. GenSim2: Scaling robot data generation with multi-modal and reasoning LLMs. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 5030–5066, 2025.

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

Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, Loic Magne, Ajay Mandlekar, Avnish Narayan, You Liang Tan, Guanzhi Wang, Jing Wang, Qi Wang, Yinzhen Xu, Xiaohui Zeng, Kaiyuan Zheng, Ruijie Zheng, Ming-Yu Liu, Luke Zettlemoyer, Dieter Fox, Jan Kautz, Scott Reed, Yuke Zhu, and Linxi Fan. DREAMGEN: Unlocking generalization in robot learning through neural trajectories. In *Proceedings of the 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 5170–5194, 2025.

Zhenyu Jiang, Yuqi Xie, Kevin Lin, Zhenjia Xu, Weikang Wan, Ajay Mandlekar, Linxi Fan, and Yuke Zhu. DexMimicGen: Automated data generation for bimanual dexterous manipulation via imitation learning. In *IEEE International Conference on Robotics and Automation*, 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):139:1–139:14, 2023.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Pete Florence, Sergey Levine, Chelsea Finn, et al. DROID: A large-scale in-the-wild robot manipulation dataset. In *Robotics: Science and Systems*, 2024. doi: 10.15607/RSS.2024.XX.120.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Zeyi Li, Yushi Yang, Shawn Xie, Kyle Xu, Tianxing Chen, Yuran Wang, Zhenhao Shen, Yan Shen, Yue Chen, Wenjun Li, Yukun Zheng, Chaorui Zhang, Siyi Lin, Fei Teng, Hongjun Yang, Ming Chen, Steve Xie, and Ruihai Wu. LeHome: A simulation environment for deformable object manipulation in household scenarios, 2026. URL https://arxiv.org/abs/2604.22363.

Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, and Gavriel State. Isaac Gym: High performance GPU-based physics simulation for robot learning. In *NeurIPS 2021 Datasets and Benchmarks Track*, 2021.

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. In *Conference on Robot Learning*, 2023.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, pp. 405–421, 2020.

Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. RoboCasa: Large-scale simulation of household tasks for generalist robots. In *Robotics: Science and Systems*, 2024. doi: 10.15607/RSS.2024.XX.050.

NVIDIA. NVIDIA Isaac Sim documentation. Online documentation, 2023. URL https://docs.isaacsim.omniverse.nvidia.com/. Accessed: 2026-05-07.

NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical AI, 2025. URL https://arxiv.org/abs/2501.03575.

Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Ajay Gupta, Abhishek Padalkar, Abraham Lee, Alex Pooley, Anirudh Gupta, Ajay Mandlekar, Ashwin Jain, Austin Tung, Alex Bewley, Alexander Herzog, Alex Irpan, Alexander Khazatsky, Aviral Rai, Avinash Gupta, Brian Wang, et al. Open X-Embodiment: Robotic learning datasets and RT-X models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

Shivansh Patel, Shraddhaa Mohan, Hanlin Mai, Unnat Jain, Svetlana Lazebnik, and Yunzhu Li. Robotic manipulation by imitating generated videos without physical demonstrations, 2025. URL https://arxiv.org/abs/2507.00990.

Deepak Pathak, Parsa Mahmoudieh, Guanghao Luo, Pulkit Agrawal, Dian Chen, Fred Shentu, Evan Shelhamer, Jitendra Malik, Alexei A. Efros, and Trevor Darrell. Zero-shot visual imitation. In *International Conference on Learning Representations*, 2018.

Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-Real transfer of robotic control with dynamics randomization. In *IEEE International Conference on Robotics and Automation*, pp. 3803–3810, 2018.

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.

SAM 3D Team, Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J. Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, Aohan Lin, Jiawei Liu, Ziqi Ma, Anushka Sagar, Bowen Song, Xiaodong Wang, Jianing Yang, Bowen Zhang, Piotr Dollár, Georgia Gkioxari, Matt Feiszli, and Jitendra Malik. SAM 3D: 3Dfy anything in images, 2025. URL https://arxiv.org/abs/2511.16624.

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4104–4113, 2016.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 23–30, 2017.

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. In *International Conference on Learning Representations*, 2024a.

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. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 51936–51983. PMLR, 2024b.

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. FoundationPose: Unified 6D pose estimation and tracking of novel objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Tianshu Wu, Jiyao Zhang, Shiqian Liang, Zhengxiao Han, and Hao Dong. Foundation feature-driven online end-effector pose estimation: A marker-free and learning-free approach, 2025. URL https://arxiv.org/abs/2503.14051.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11097–11107, 2020.

Mengda Xu, Han Zhang, Yifan Hou, Zhenjia Xu, Linxi Fan, Manuela Veloso, and Shuran Song. Dexumi: Using human hand as the universal manipulation interface for dexterous manipulation. *arXiv preprint arXiv:2505.21864*, 2025.

Zhengrong Xue, Shuying Deng, Zhenyang Chen, Yixuan Wang, Zhecheng Yuan, and Huazhe Xu. DemoGen: Synthetic demonstration generation for data-efficient visuomotor policy learning. In *Robotics: Science and Systems*, 2025.

Sherry Yang, Yilun Du, Seyed Kamyar Seyed Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *International Conference on Learning Representations*, 2024.

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

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag Sanketi, Michael Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Kuang-Huei Lee, Isabel Leal, Yuheng Kuang, Ryan Julian, Nikhil Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Pete Florence, Chelsea Finn, Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: Vision-Language-Action models transfer web knowledge to robotic control. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183. PMLR, 2023.

A LIMITATION

While Grounded-Dreamer substantially automates data generation, it currently requires a minimal, one-time collection of reference images—typically one scene overview and 1–4 object perspectives—to initialize the environment. Given that image-conditioned editing models are still rapidly evolving, the success rate for our synthesized initial frames is approximately 65%. However, we view this as a robust baseline that will naturally improve as foundational generative models mature.

While our experiments establish the framework's effectiveness with hundreds of trajectories per task, the scale was primarily constrained by available computational resources rather than algorithmic bottlenecks. The inherent design of our pipeline is highly scalable; as hardware constraints are alleviated, we anticipate extending this paradigm to massive-scale data generation to further unlock the scaling potential of generalist robot policies.

B BROADER IMPACTS

Our work introduces a scalable pipeline for generating high-fidelity, verified synthetic data, which has significant implications for the future of Embodied AI. By reducing the reliance on costly and labor-intensive real-world data collection, our approach facilitates the scaling of robotic learning across diverse tasks and environments. While this may accelerate the deployment of autonomous systems, it also underscores the importance of addressing potential biases in world models and ensuring the safety of policies trained on synthetic data. We believe this research promotes more accessible and efficient robot learning, though we encourage practitioners to maintain rigorous safety protocols when deploying such models in real-world human-centric environments.

C ADDITIONAL QUALITATIVE VISUALIZATIONS

In this section, we provide extended visualization results to facilitate a comprehensive qualitative analysis of the *Grounded-Dreamer* framework.

C.1 AUTOMATED SIMULATION-BASED DATA ACQUISITION

Figure 8 illustrates the automated data collection process within the simulation environment. For each task, the top row displays the original simulation frames, while the bottom row shows the frames after applying our

C.2 VIDEO ROLLOUTS FROM GROUNDED-DREAMER

We present additional video rollout results generated by our world model in Figure 9. These sequences demonstrate the *Grounded-Dreamer*'s capability to predict physically plausible future states.

D DETAILS OF INITIAL FRAME SYNTHESIS

We employ *Nano Banana 2* (Google DeepMind, 2026) as the generative backbone for synthesizing initial frames that are distributionally aligned with the real world. The implementation details are as follows:

- **Input Modalities:** The synthesis process takes two distinct inputs: (1) a composite image consisting of simulated object assets overlaid on a real-world scene, and (2) a reference photograph containing the actual objects to provide essential structural and illumination constraints.
- **Prompting Technique:** We utilize a structured prompt to guide the generative harmonization: "There is a simulated object in Image 1, identified as [object]. Image 2 displays its true, real-life form. Please substitute the simulated object in Image 1 with a new version in a realistic style, while keeping the position and orientation of the original simulated object from Image 1. The replacement should not be aligned with the realistic object in Image 2.



Figure 8: Additional visualizations of automated data collection in simulation.

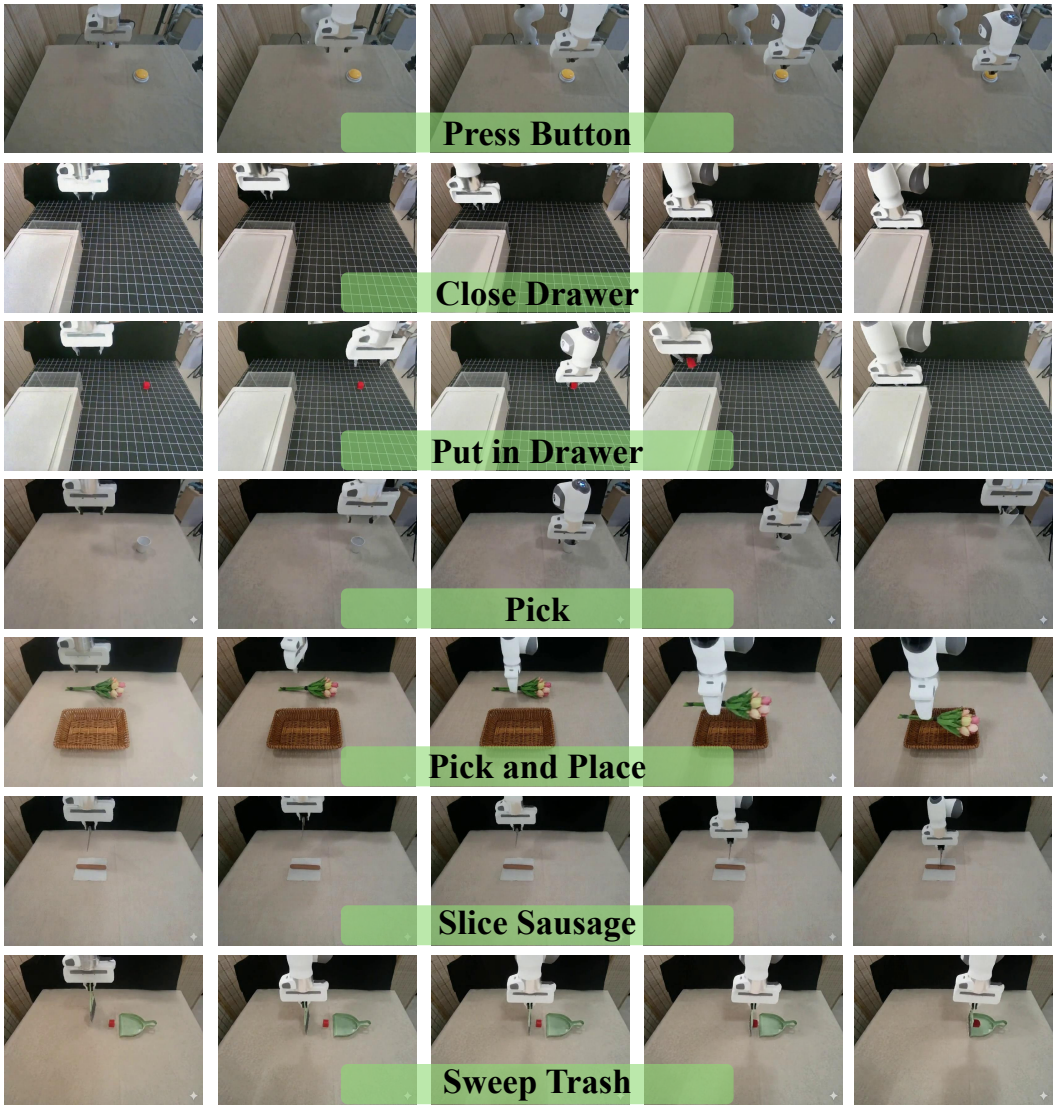

Figure 9: Additional visualizations of Grounded-Dreamer video generation.

The real object serves only as a visual guide for appearance. The background must not be altered."



Figure 10: Qualitative Visualizations of Refined Initial Frames.

Figure 10 provides qualitative visualizations of these synthesized frames, illustrating the seamless integration of virtual assets into diverse physical contexts.

E Experimental Details

In this section, we provide comprehensive details regarding the training and inference pipelines of our framework to ensure reproducibility.

E.1 Fine-Tuning Details of Cosmos

We fine-tune the 2B-parameter Cosmos Predict2.5 Video2World model (NVIDIA et al., 2025) as the generative backbone. The model is optimized using a learning rate of 4.315×10^{-5} on a cluster of eight NVIDIA A800 GPUs with context parallelism. Training is conducted for 1,400 to 2,000 iterations with a global batch size of 32. We use a temporal length of 125 frames for the *Put in Drawer* task and 93 frames for all other tasks. The fine-tuning process requires 74 GB of VRAM per GPU.

E.2 Training Details of the Inverse Dynamics Model (IDM)

The IDM (Jang et al., 2025) features a SigLIP2-Large encoder and a flow-matching action head. It is trained across the eight A800 GPUs for 5,000 to 10,000 steps with a global batch size of 512 and a 10^{-4} learning rate. Training is performed in BF16 mixed precision to ensure numerical stability.

E.3 Synthetic Data Generation and Action Extraction

All inference and data generation tasks are also executed on the eight NVIDIA A800 GPU cluster.**Video Generation via Cosmos.** We generate rollout sequences conditioned on five replicated initial frames. For the *Put in Drawer* task, the model produces 125-frame sequences, while 93-frame sequences are generated for other tasks. The inference memory footprint is approximately 16 GB per GPU.**Action Extraction via IDM.** Robot actions are extracted from the generated sequences using the IDM, which demonstrates high throughput (0.5 seconds per trajectory). This process is memory-efficient, requiring only 4 GB of VRAM per GPU. The average task success rate during this generation phase is approximately 45