2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Physically-Grounded Data Generation via Video Diffusion Models

Sriram Yenamandra[1], Dorsa Sadigh[1]

Abstract— **Existing datasets for training generalist manipulation policies often lack diversity in object variety and initial states, limiting the range of physically grounded interactions present in them. Consequently, these policies struggle with unseen object shapes, sizes, or unfamiliar object poses. Manually collecting real-world trajectories with diverse physical interactions is tedious, time-consuming, and expensive, underscoring the need to generate these autonomously. Simulators offer a scalable pathway to autonomously generate trajectories by enabling extensive variation not only in tasks (e.g., objects, object properties, and initial conditions), but also in the robot behaviors required to solve these tasks. We develop a data generation pipeline that autonomously produces physically grounded trajectories in simulation using video diffusion models. Our approach first simulates random initial conditions across various tasks using a diverse asset library. A video diffusion model generates videos of a robot performing these tasks in physically diverse scenarios, which are then fed to a learned goal-conditioned planner to extract actions that closely follow the generated videos. Unlike prior trajectory generation methods, our pipeline generalizes to new objects across multiple tasks without relying on human demonstrations. Using our approach, we generate a simulation dataset PHYSVIVID, containing 5k+ demonstrations involving 400+ objects. We demonstrate the effectiveness of PHYSVIVID by fine-tuning robot policies on it, and demonstrating generalization of policies to unseen objects with varying shapes, textures, and sizes, as well as to unseen object categories. See videos on our website: https://sites.google.com/view/physvivid/.**

I. INTRODUCTION

Inspired by the remarkable success of multimodal large language models that can solve a wide variety of text- and image-based tasks [1], [2], [3], [4], [5], recent research has attempted to follow the same recipe to build generalist robot policies [6], [7], [8], [9], [10], [11], [12]. However, unlike text or image tasks – where abundant and diverse data is freely available online – existing datasets for training generalist manipulation policies are limited in terms of scale and diversity essential for training a generalist policy [1], [13], [14]. This lack of variety, in particular lack of diversity in object properties or *physical* interactions in the world leads to poor performance of robot policies trained on these datasets when encountering unfamiliar objects that require new physically-aware interactions [15]. As an example, the BridgeData V2 [13] dataset mainly contains manipulation data of a small set of plastic cups (Fig. 1; left). A policy trained on this dataset often struggles with picking up larger or heavier cups – a behavior that is seemingly similar to picking up smaller cups but can require a completely different motion, torque, or grasp.

Manually collecting real-world data with sufficient diversity of physical interactions is tedious, time-consuming, and expensive, often requiring experts to ensure high quality. Consequently, there is a clear need to collect data autonomously, but real-world attempts are hindered by inherent limitations in object variety, manual resets, and lack of reliable success detection mechanisms [16], [17]. In contrast, simulators have the potential to provide a scalable way to autonomously generate new trajectories with controlled quality and diversity. They enable extensive variation not only in the tasks (e.g., different objects, object properties, and scene setups) but also in the behavior of the robot when performing the task with such object variations (e.g., different grasps needed to lift small or large cups as in Fig. 1).

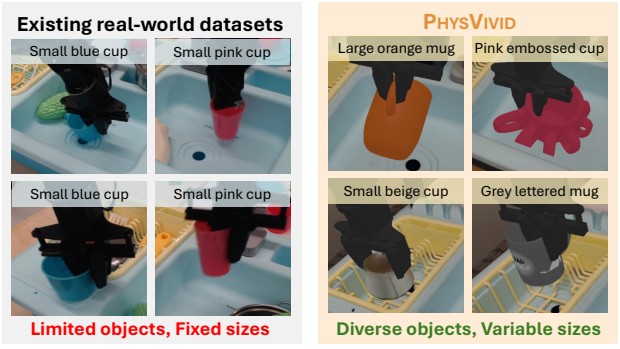

Fig. 1: **Diversity of physically grounded motions in PHYSVIVID.** Left shows samples from BridgeData V2 [13] with limited object diversity: a small set of plastic cups get used throughout the dataset. Right shows samples from our generated dataset **PHYSVIVID**, which expands the diversity of physical interactions using a rich asset library consisting of 400+ objects.

To scale data generation in simulation, previous work has proposed using digital twins — virtual replicas of real-world environments [18], [19]. However, most methods utilizing digital twins focus on randomizing static variations in the scene as opposed to autonomously *expanding the diversity of physically grounded interactions and motions*. Most popular methods for trajectory generation in simulation *rely on extrapolating human demonstrations*, which restricts motions to previously observed contact points [20], [21], [22]. Such methods also often fail to handle novel tasks or unseen objects that demand new affordances. For instance, if the provided human demonstrations primarily grasp small cups with robot grippers positioned on cup's exteriors, these generation methods would fail to effectively grasp wider cups that require different gripping strategies, e.g., grasping the rim of the cup (Fig. 1). This highlights a critical research gap: *how can we* ***increase diversity of physical interactions*** *in simulation* ***without relying on extensive human input****, thereby ensuring that the robot can generalize to a wide range of real-world situations?*

[1]Stanford University
Correspondence to ysriram@stanford.edu

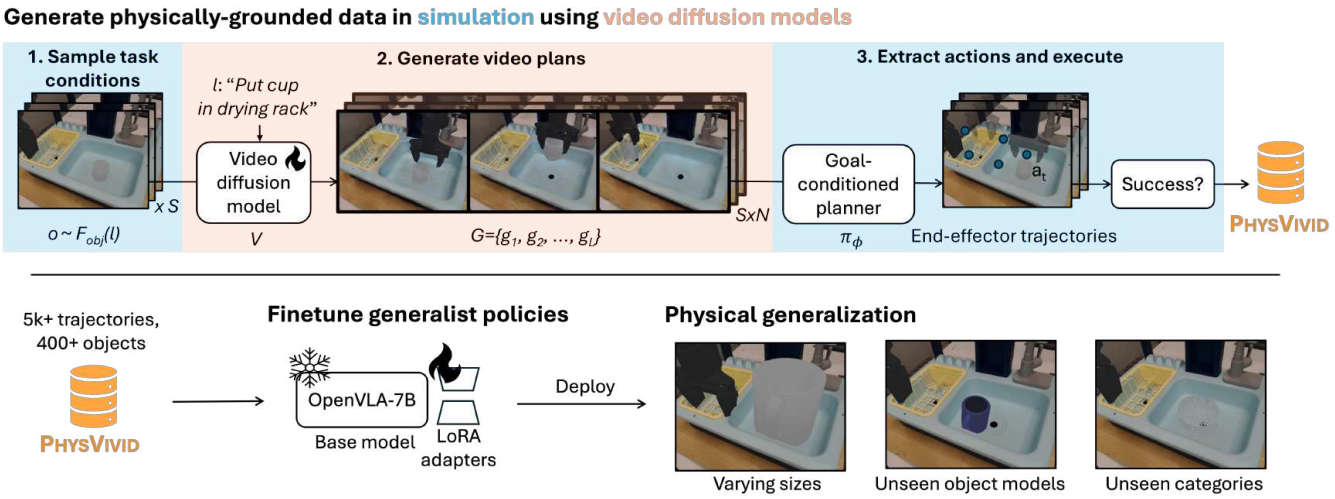

Fig. 2: **Overview of our generation pipeline.** We generate physically grounded trajectories in simulation using the motion generation capabilities of video diffusion models. We first simulate diverse initial conditions for each task (e.g., "Put cup in drying rack") using a rich object library. A video diffusion model (V) generates N videos of robot solving the task. A goal-conditioned planner (π_ϕ) extracts actions from generated videos, which are rolled out in simulation. The resulting trajectories are filtered based on task success to create our **PHYSVIVID** dataset comprising of 5k+ trajectories of 400+ objects. Training generalist policies [8] on **PHYSVIVID** improves their generalization to objects of varying shapes, textures, and sizes, as well as to objects from unseen object categories.

To address these challenges, we combine the strengths of simulators and video diffusion models. Simulators provide necessary physical grounding while allowing for extensive variation of objects and their physical properties. Meanwhile, we leverage the motion generation capabilities of video diffusion models, pretrained on video datasets that encompass a wide range of physical interactions with diverse objects, for autonomous generation at scale. We use these models to sample diverse video demonstrations of robots performing physical tasks, surpassing the motion diversity achievable by prior generation methods. Then, we leverage a goal-conditioned planner that closely follows the generated video, frame-by-frame, within a simulated environment to produce the corresponding actions required for achieving the robot's behavior in the new physical conditions of the task. This synergy between video diffusion models and simulations enables autonomously producing diverse physical interactions – overcoming the limitations of previous data generation methods that would not benefit from task coverage, pixel-level guidance, and motion diversity of pretrained video generation models.

Our method leverages off-the-shelf video diffusion models [23] fine-tuned on a small amount of scripted data to generate diverse videos of manipulation policies. We then simulate diverse initial conditions using an asset library of more than 400+ object models. In this work, we focus on generalization to objects with diverse shapes and sizes as our main instance of physical grounding. Variations to these properties ensure that the robot behaviors generated are grounded in real-world physics, capturing a wide range of realistic and diverse interactions. Our diffusion model generates robot videos under these varied conditions. Using a goal-conditioned behavior cloning policy [24], we generate actions in simulation that closely follow the generated video in diverse physical variations of the scene (for example, the actions needed to pick different cups in Fig. 1). We demonstrate that our generated physically-grounded dataset, which we call **PHYSVIVID** (Physically grounded data via Video Diffusion), enables generalization to unseen objects and object variations. Fine-tuning a pretrained policy such as OpenVLA [8] on this grounded data improves its performance by an average of 15% across three tasks evaluating physical generalization to objects with varying shapes, textures, and sizes, as well as to unseen object categories.

We summarize our contributions below:

- We develop a pipeline to autonomously generate *physically grounded trajectories* in simulation by leveraging a video diffusion model and a goal-conditioned planner.
- Using our pipeline, we generate a dataset, **PHYSVIVID**, featuring 5k+ trajectories, manipulating 400+ objects, and showcasing a diverse range of physical motions.
- We demonstrate that fine-tuning generalist robot policies on **PHYSVIVID** improves generalization to unseen objects with varying shapes, textures, and sizes, and from unseen object categories by an average of 15% across three simulated tasks.

II. Related Work

In this section, we discuss the shortcomings of datasets used currently for training generalist policies, review previous methods for trajectory generation in simulation, and explore the use of image diffusion models in prior manipulation works.

Datasets for generalist manipulation policies. Generalist robot manipulation policies [6], [25], [9], [7] are typically trained using data collected through robot teleoperation, which is difficult to scale. Recent efforts have sought to expand training datasets using human videos [26], [27], [28], simulations [20], [29], and cross-embodiment data [30], [31], [32]. Although promising, integrating videos and cross-embodiment data often requires spe-

cialized solutions to bridge the embodiment gaps between robots and humans. In contrast, our work demonstrates the potential of using simulations to expand datasets. The simulated data used in prior works [20], [29], [11], [33] tends to be limited in scale and diversity, often relying on collections of existing small-scale simulation datasets. In this work, we provide a pipeline for scaling the amount of physically-grounded robot data in simulation.

Trajectory generation in simulation. Prior works for autonomous trajectory generation in simulation rely on skills derived from reinforcement learning [34], [35], [36], [37] or grasp and motion planning [38], [39] — at times using large language models for planning [38]. However, reinforcement learning is challenging to tune and often requires task-specific training. Domain randomization under these setups can increase environmental diversity, but behavioral diversity must still be discovered through reinforcement learning or planning, increasing learning complexity. Grasp samplers, on the other hand, struggle with objects of varying sizes or complex motions like flipping of cups. Other approaches [20], [21], [22], [19] attempt to mimic object contact points from a few human demonstrations, yet fail to generalize to new objects with varying shapes and sizes that demand different affordances. In contrast, we employ video diffusion models [40], [23], which can generalize to unseen objects and generate useful robot trajectories across multiple tasks, without relying on human demonstrations.

Video generation for robot manipulation. Numerous works have explored using image and video diffusion models [24], [40], [41], [42], [43], [44], [45] for visual planning. These approaches typically employ a diffusion model to generate visual plans of robot or humans [46] executing tasks, paired with an inverse dynamics model to extract the underlying low-level actions. However, the real world deployment of this pipeline achieves only modest success rates [17]. We instead re-purpose this pipeline for generating diverse physical interactions to train generalist policies via imitation learning. We execute the trajectories in simulation, verify task completion, and thereby account for potential failures from our generation pipeline. Although recent works [17] have similarly attempted to use image diffusion models for autonomous data collection on real robots, they are constrained by the inherent limited diversity of objects accessible in the real world and the shortcomings of automated success detection mechanisms [16]. Concurrent work [47] investigates video-generation-based imagination of future robot behaviors, but unlike us, does not perform simulator-based verification or target physical diversity. Additionally, recent work has attempted to extract supervision from generated videos for simulator-based trajectory optimization [48] or real-robot retargeting [49], but again do not consider varying physical properties. Another line of works augment existing datasets by using image diffusion models to generate visual variations [50], [51] or slightly off-trajectory images [52]. In contrast, our approach generates entirely new interactions with diverse objects, with the aim of enriching the physical diversity of datasets.

III. Physically-Grounded Data Generation

This section details our data generation pipeline for autonomously producing diverse, physically-grounded robot trajectories in simulation. These trajectories involve interactions with varying objects of different shapes and sizes across multiple tabletop manipulation tasks. To generate these, we integrate the motion generation capabilities of video diffusion models with scalable physics-based simulation. Concretely, our method first uses a simulator augmented with a rich object library to create diverse conditions for different tasks (Section III-A). We then leverage diffusion models for generating videos of robots performing these physically diverse tasks (Section III-B). We then use a goal-conditioned planner for grounding the videos and generate low-level actions that follow the videos frame-by-frame (Section III-C). Additionally, we use simulators to validate the effectiveness of the generated actions for solving the desired tasks (Section III-D).

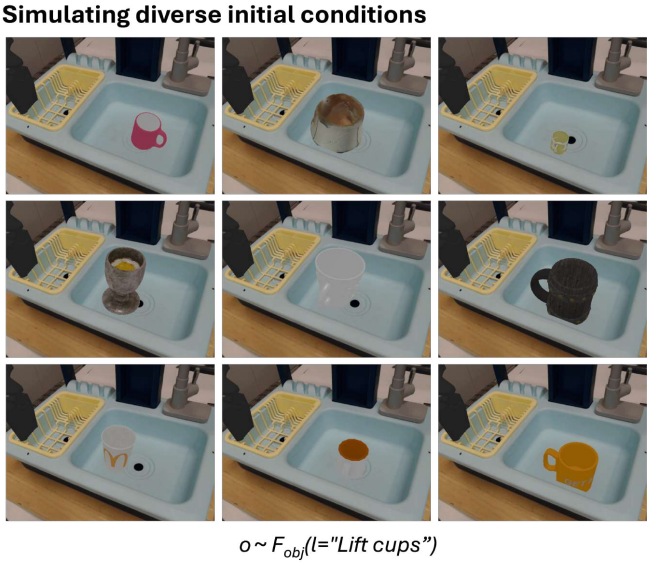

Fig. 3: **Simulation of diverse initial conditions.** We generate diverse initial conditions for each task (e.g., *Lift cup*) by sampling object models from a rich object library of 400+ objects and sampling appropriate scales and poses for them.

A. Simulation of Diverse Initial Conditions

Recognizing the potential of simulations to scale autonomous data generation [53], [54], we firstly create a diverse set of initial environment states in simulation. We pool 3D object models from existing object datasets to create an object library $\mathcal{O}$. The use of a large collection of object models allows the generation of trajectories for objects with different shapes and textures – significantly surpassing the diversity of objects accessible in real-world environments. For instance, the 3D object datasets provide as many as 500 different instances of cups and glasses (illustrated in Fig. 3). We simulate a diverse set of initial conditions by randomizing these object assets, in addition to their scales and poses.

Formally, let $T = l_1, l_2, ..., l_n$ represent a set of language instructions (e.g., *Lift cup*), each corresponding to a specific task. Define $F_{obj} : T \to 2^{\mathcal{O}}$ as a function that maps each instruction $l \in T$ to a subset of object models suitable for that task. Our generation procedure operates as follows: for each instruction $l \in T$, we first sample an object model $o \in F_{obj}(l)$. Before

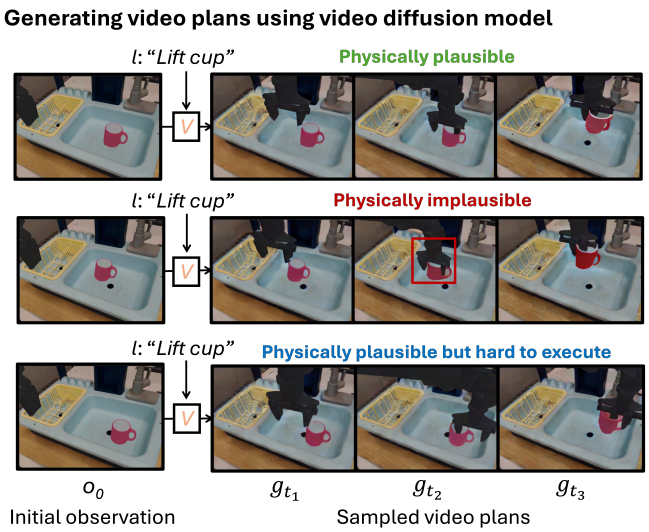

Fig. 4: **Generating multiple video plans using a video diffusion model.** The images on the left illustrate representative initial states for the *Lift Cup* task, while the frames $(g_{t_1}, g_{t_2}, g_{t_3})$ on the right display video plans produced by our model V, subsampled at timesteps t_1, t_2, and t_3. The first row shows a physically plausible scenario of the robot lifting a cup. The second row depicts an implausible scenario where the robot's gripper intersects the cup's edges. The third row shows a sequence that requires precise positioning to grip the cup's handle.

positioning the sampled object model o in the environment, we assign it appropriate scale and pose by sampling values from feasible ranges that are determined based on task l and geometry of o. For example, for the *Lift cup* task illustrated in Fig. 3, the scales and poses of the cup objects are sampled so that the cups can fit within the sink's bounds. We position the object in the environment to completely instantiate the task, and repeat this procedure to sample multiple initial task conditions (like the ones illustrated in Fig. 3). With this systematic randomization of initial task conditions, we enhance the physical diversity of objects in our generated dataset for learning generalist policies.

B. Generating Robot Plans via Video Diffusion Model

Simulators can scale up object and visual diversity, but this by itself does not give us the training supervision necessary to train generalist policies – the robot movements that would solve different tasks. Generating robot movements across different manipulation tasks, objects and object properties requires reasoning about the necessary task-specific affordances and motions. For example, lifting cups with different shapes and sizes in Fig. 1 would need different gripping strategies. Our key insight is that video diffusion models, typically pretrained on large video datasets encompassing different objects and tasks, possess this physical understanding.

Video diffusion models are improving rapidly, achieving increasingly realistic and coherent generations, we find that open-source video diffusion models still fail to generate videos of robots executing tasks, as these are primarily trained on human videos. So, we customize our video diffusion model to the target robot embodiment and environment by finetuning it on a small dataset of scripted trajectories generated in simulation using a subset of objects $O_s \in \mathcal{O}$.

We leverage our customized video diffusion model to generate videos that show different ways in which a robot could complete a particular task starting from a given initial state (Fig. 4 top row). Although some sampled videos may be implausible (Fig. 4 middle row) or challenging to execute (Fig. 4 last row), producing a diverse set of execution trajectories—such as varying gripping techniques—enhances the likelihood that at least one trajectory will succeed. Precisely, given a language instruction l and the initial third-person observation image I, we use a video diffusion model V to generate N videos, $G_1, G_2, \ldots G_N$, of the robot solving the task. Out of these, the video plans that cannot result in task-completing trajectories get filtered out in subsequent stages (Section III-D). We generate video plans for different tasks and sampled initial task conditions. Next, we describe the process of extracting low-level actions from these generated videos.

C. Extracting Low-level Actions from Videos

While video diffusion models can generate visual demonstrations of robot task executions, training generalist policies through imitation learning requires low-level actions that robots can physically execute. To bridge this gap, we train a goal-conditioned behavior cloning policy, π_ϕ parameterized by ϕ, which outputs the actions (Fig. 5 bottom) that closely follow the motions depicted in the generated videos (Fig. 5 top). Given a generated video plan G consisting of L frames: $\{g_1, g_2 \ldots, g_L\}$, the policy sequentially attempts reaching the environment state depicted in each subgoal frame. Specifically, for a current observation o_t and a subgoal frame $g \in G$ it predicts a chunk of k next actions to reach the subgoal g: $a_{t:t+k} = \pi_\phi(o_t, g)$.

We train this policy using the same scripted trajectories on the object subset O_s that were used for customizing the video diffusion model in Section III-B. At the time of data generation, we sample a new subgoal every m timesteps, and execute the first $e < k$ actions from the predicted chunk. In Fig. 5, for the sake of simplicity we illustrate this procedure for $e = m$, i.e., a new subgoal is sampled every e timesteps, and the first e actions of the predicted chunk are executed in simulation. The figure shows a rollout from π_ϕ in simulation (bottom sequence), that attempts to follow the generated video (top sequence).

The video diffusion model may occasionally generate physically implausible frames (Fig. 4). However, the goal-conditioned planner is trained only on physically valid simulation trajectories and therefore tends to produce feasible motions from current state even when subgoals are imperfect. Furthermore, trajectories are executed and validated in simulation, and failures are filtered using success detection, which we describe next.

D. Success Detection and Filtering

The pipeline discussed thus far is capable of generating trajectories that attempt to solve the task under diverse simulated initial conditions. However, before training on these trajectories, it is essential to validate their success at accomplishing the intended tasks. This validation is particularly important when

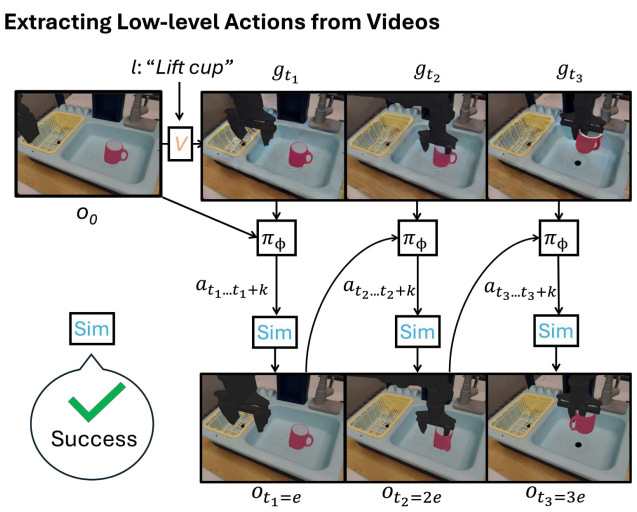

Fig. 5: **Extracting actions from generated videos.** The top row shows a video of the robot executing *Lift cup* task from our video diffusion model. Our goal-conditioned planner extracts low-level actions from this video. The bottom row shows the corresponding robot trajectory executed in simulation using the extracted actions. In this illustration, first $e < k$ actions from the predicted chunk are executed before the next subgoal is sampled. A success detection mechanism determines if the robot succeeds.

generating data for manipulating uncommon objects or objects with atypical sizes, where execution failures are more likely. To ensure quality of the generated data, after executing the extracted plans in simulator, we apply task-specific success detection logic that uses privileged information from simulator (e.g., object poses) to determine whether a trajectory successfully completes the task. For example, for the *Lift cup* task in Fig. 5, the success detection logic would mark the episode as successful if the cup is raised above a certain height with the robot holding on to it. We code up similar success criteria for other tasks. While this requires task-specific success definitions, in simulation these are typically simple geometric checks using privileged state (e.g., object height or location). For more complex tasks, learned success classifiers can be used. Using these criteria, trajectories that fail to meet the success conditions are filtered out, giving a set of physically grounded trajectories suitable for training generalist policies – which we call the **PHYSVIVID** dataset.

IV. PHYSVIVID

In this section, we discuss details of **PHYSVIVID**, the physically-grounded dataset we generate using the pipeline described above. We go over the details of the simulated environment and the tasks in our dataset.

Simulated environment. For simulating diverse initial conditions, we utilize a simulated environment from SimplerEnv [55], which closely mirrors a real environment from the BridgeData V2 dataset [13]. Specifically, we use a digital twin of a toy sink environment that is part of pretraining mixture of generalist policies [8], and also commonly used for evaluating them. Our asset library comprises of over 400 object models of cups, sourced from multiple object datasets [56], [57], [58], [59], [60]. We use WidowX 250 arm as our robot embodiment and a mounted third person camera is used to collect observations.

Tasks in PHYSVIVID. In total, we generate 5508 trajectories across the following three tasks (Fig. 6) using 427 object models:

- **`lift`**: The task is to lift an object placed in the sink.
- **`put_in_rack`**: The toy sink environment has two compartments: a sink and a smaller drying rack. The task involves picking the referred object and placing it in sink.
- **`put_in_sink`**: The task is to pick an object from drying rack and place it in sink. This is slightly more challenging than the prior task as it requires finding grasps in a restricted space.

V. EXPERIMENTS

In this section, we describe the experiments we conduct to test the effectiveness of our data generation pipeline. We primarily attempt to answer the following questions: can finetuning pretrained policies on our generated data improve their physical generalization to i) varying object shapes, textures and sizes, and ii) unseen object categories?

In Section V-A, we describe the evaluation settings used to test physical generalization. We then present our results for physical generalization (Section V-B) and sim2real transfer (Section V-C).

A. Evaluation settings

We use SimplerEnv [55] to conduct comprehensive evaluation of generalist policies focusing on their physical generalization capabilities. In particular, we study their ability to generalize to objects of varying *shapes*, *sizes* and *textures*, and objects from unseen *categories*. The settings we study are:

- **Unseen Object Models from Seen Category**: In this setting, new object models belonging to object categories present in **PHYSVIVID** (e.g., cups) are used for evaluation. Although these object categories are already present in our dataset, the new object models feature entirely novel shapes, textures, or sizes (like the examples in top row of Fig. 6). We create 284 scenarios per task using 50 new object models.
- **Unseen Object Category**: Objects belonging to an entirely new object category (e.g., bowls) are introduced to further assess the policy's generalization capabilities. Different categories of objects have entirely different shapes (examples in bottom row of Fig. 6), requiring novel grasping strategies (for example, bowls are typically wider than cups). We create 254 scenarios per task using 43 object models.

B. Results for physical generalization in simulation

We evaluate our model against two main baseline policies: OpenVLA [8] and Octo [7]. These models are pretrained on 970k and 800k episodes respectively from the Open-X Embodiment dataset [1] and have demonstrated zero-shot performance on Bridge-like tasks like the ones shown in Fig. 6. We then fine-tune an OpenVLA policy on **PHYSVIVID**, which we call OpenVLA-FT (ours), and compare its performance on physical generalization with OpenVLA and Octo out-of-the-box (see Fig. 6). We conduct evaluations on the evaluation settings discussed in Section V-A. Our evaluation uses 538 diverse scenarios for each task, involving 93 previously unseen objects, each

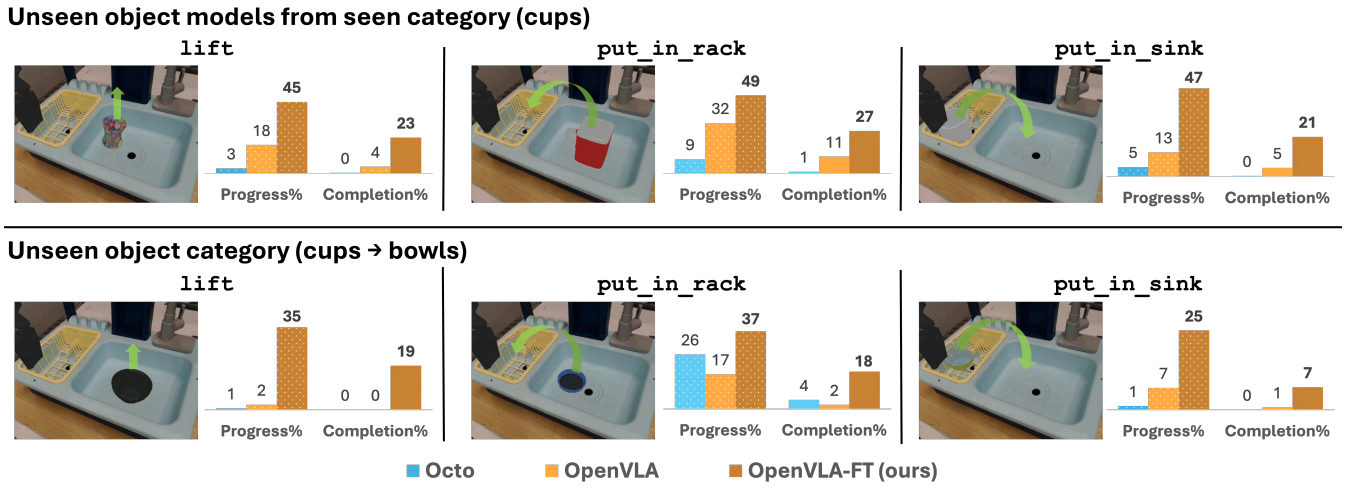

Fig. 6: **Physical generalization results.** To test the effectiveness of models finetuned on our **PHYSVIVID**, we evaluate their ability to generalize to unseen object models from seen categories (cups; top row) and unseen object categories (bowls; bottom row). Alongside task completion rates (solid bars), we report task progress rates (patterned bars), which assign partial scores for successfully grasping the target object. Finetuning OpenVLA [8] on **PHYSVIVID** (OpenVLA-FT) significantly improves task completion rates and task progress rates on both seen and unseen object categories across all tasks.

with a unique texture and shape. The objects are randomly scaled across episodes resulting in a height variation of 4cm-12cm. [1]

We find that generalist policies like Octo [7] and OpenVLA [8] achieve task completion rates under 11% across all tasks. These policies are typically trained on real-world datasets having a fixed set of objects (like small plastic cups) – causing them to underperform on novel objects with diverse shapes, sizes and textures, used in our evaluation settings. On finetuning OpenVLA [8] on our physically-grounded data containing 5k+ trajectories of 400+ objects, we observe absolute improvements of $16 - 18\%$ for seen categories (cups) and $6 - 19\%$ for unseen categories (bowls) in task completion rates. These results demonstrate that our data generation pipeline by covering a wide range of physical interactions, can improve the generalization of policies to unseen objects with diverse textures, shapes, and sizes.

C. Evaluations on real robot

We validate that the data generated in the simulation can in fact be useful for real-world execution, i.e., the policies trained on **PHYSVIVID** transfer reliably to a real WidowX 250 arm. We conduct experiments for `lift` task with new cups (Fig. 7). Our results demonstrate positive sim2real transfer – OpenVLA, when fine-tuned **solely** on simulation data, achieves a success rate of 70% across 20 trials (Table I) outperforming base OpenVLA [8]. We note that we have not used any real-world demonstrations from our setup, and the benefits solely arise from the physically grounded data generated in simulation.

D. Analysis of data generation pipeline

We demonstrate robustness of our data generation pipeline to noisy generations from video diffusion model and discuss the amount of data needed to customize video diffusion model.

[1]Additional visualizations of the diverse evaluation scenarios and their corresponding rollouts are available on our website: https://sites.google.com/view/physvivid/.

Model	OpenVLA	OpenVLA-FT
`lift`	4/20	14/20

TABLE I: **Real robot evaluations.** OpenVLA improves at lifting unseen cups after finetuning on simulated data.

Fig. 7: **Successful lifts on real robot.** OpenVLA-FT succeeds at lifting cups of varied shapes, textures and sizes.

Robustness to noisy generations. The videos generated by the video diffusion model occasionally exhibit visual inconsistencies, such as changes in object shape or color midway through a trajectory (Fig. 8). However, we find that our goal-conditioned planner remains relatively robust to these imperfections. Nevertheless, the planner still depends on the generated subgoals to capture task-relevant cues – like grasps as in Fig. 8.

Amount of scripted data. To customize off-the-shelf video diffusion models, we use a small set of scripted trajectories. The generation success rate, evaluated across 150 scenarios spanning three tasks, remains stable even if the number of objects in the finetuning data decreases from 25 to 5, dropping only slightly from 44% to 41% – implying that trajectories on as few as 5 objects suffice.

VI. DISCUSSION

Summary. Existing real-world manipulation datasets tend to be limited in the diversity of physical interactions, often manipulating a limited set of objects. This work proposes

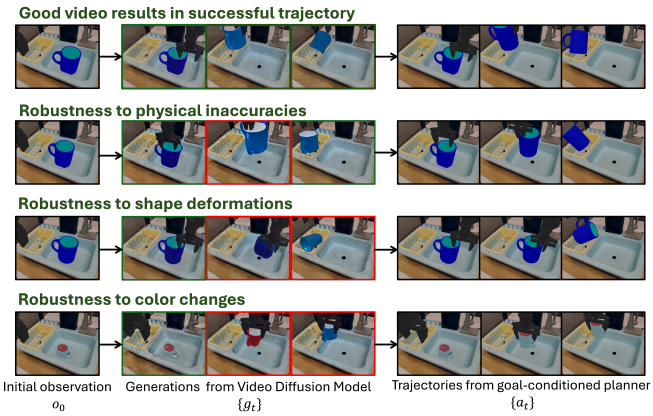

Fig. 8: **Planner's robustness to noisy video generations.** Examples of the "Put cup in drying rack" task.

an autonomous data generation pipeline for expanding the diversity of physical interactions. The proposed pipeline combines motion generation capabilities of video diffusion models with physics-based simulations. A simulator augmented with a diverse object library is used to generate diverse initial conditions. A video diffusion model generates video plans which are rolled out in simulation using a goal-conditioned planner. Using our pipeline, we generate a dataset **PHYSVIVID** of 5k+ physically-grounded trajectories manipulating 400+ objects. Finetuning generalist manipulation policies on **PHYSVIVID** improves their generalization to unseen objects with varying shapes, textures, and sizes, as well as unseen object categories.

Limitations. Our work highlights the potential of video generation models for autonomous data generation; however, the scale of our dataset and experiments remain limited, preventing full realization of our proposed pipeline's capabilities. Future work could extend our approach to generate data across a broader range of tasks, environments, physical properties, and object categories. Our experiments primarily focus on pick-and-place tasks – extending this approach to more dexterous tasks may be challenging, as obtaining a high-quality goal-conditioned planner for such tasks is non-trivial. Our current experiments focus on short-horizon pick-and-place tasks. Extending this plan-then-execute framework to long-horizon or high-precision assembly tasks may be challenging due to compounding errors and may require periodic replanning, which we leave for future work. Additionally, we finetune a video diffusion model to produce physically realistic robot task executions. Moving forward, we hope that rapid advancements in this space would allow us to use these models directly off-the-shelf. We believe our pipeline holds substantial potential to scale and diversify datasets for training generalist policies, and this work represents a step toward realizing that vision.

VII. ACKNOWLEDGMENTS

We thank members of the ILIAD lab, namely, Jensen Gao, Amber Xie and Shuang Li for feedback on the draft. This work was supported by ONR project N00014-22-1-2293, NSF 2132847 and ONR N00014-25-1-2479.

REFERENCES

[1] A. O'Neill, A. Rehman, A. Gupta, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models," *arXiv preprint arXiv:2310.08864*, 2023.

[2] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[3] M. Deitke, C. Clark, S. Lee, R. Tripathi, Y. Yang, J. S. Park, M. Salehi, N. Muennighoff, K. Lo, L. Soldaini *et al.*, "Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models," *arXiv preprint arXiv:2409.17146*, 2024.

[4] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[5] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[6] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski *et al.*, "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control," in *Conference on Robot Learning (CoRL)*, 2023.

[7] Octo Model Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, C. Xu, J. Luo, T. Kreiman, Y. Tan, L. Y. Chen, P. Sanketi, Q. Vuong, T. Xiao, D. Sadigh, C. Finn, and S. Levine, "Octo: An open-source generalist robot policy," in *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.

[8] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. P. Foster, P. R. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn, "OpenVLA: An open-source vision-language-action model," in *8th Annual Conference on Robot Learning*, 2024.

[9] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, L. X. Shi, J. Tanner, Q. Vuong, A. Walling, H. Wang, and U. Zhilinsky, "π_0: A vision-language-action flow model for general robot control," 2024.

[10] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai *et al.*, "$\pi_{0.5}$: a vision-language-action model with open-world generalization," in *Conference on Robot Learning (CoRL)*, 2025.

[11] L. Wang, X. Chen, J. Zhao, and K. He, "Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[12] G. R. Team, S. Abeyruwan, J. Ainslie, J.-B. Alayrac, M. G. Arenas, T. Armstrong, A. Balakrishna, R. Baruch, M. Bauza, M. Blokzijl, S. Bohez, K. Bousmalis *et al.*, "Gemini robotics: Bringing ai into the physical world," 2025. [Online]. Available: https://arxiv.org/abs/2503.20020

[13] H. R. Walke, K. Black, T. Z. Zhao, Q. Vuong, C. Zheng, P. Hansen-Estruch, A. W. He, V. Myers, M. J. Kim, M. Du *et al.*, "Bridgedata v2: A dataset for robot learning at scale," in *Conference on Robot Learning*. PMLR, 2023, pp. 1723–1736.

[14] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis *et al.*, "DROID: A Large-Scale In-the-Wild Robot Manipulation Dataset," in *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[15] J. Gao, S. Belkhale, S. Dasari, A. Balakrishna, D. Shah, and D. Sadigh, "A taxonomy for evaluating generalist robot manipulation policies," *IEEE Robotics and Automation Letters (RA-L)*, 2026.

[16] S. Mirchandani, S. Belkhale, J. Hejna, E. Choi, M. S. Islam, and D. Sadigh, "So you think you can scale up autonomous robot data collection?" in *8th Annual Conference on Robot Learning*, 2024.

[17] Z. Zhou, P. Atreya, A. Lee, H. R. Walke, O. Mees, and S. Levine, "Autonomous improvement of instruction following skills via foundation models," in *8th Annual Conference on Robot Learning*, 2024.

[18] M. Torne, A. Simeonov, Z. Li, A. Chan, T. Chen, A. Gupta, and P. Agrawal, "Reconciling reality through simulation: A real-to-sim-to-real approach for robust manipulation," in *Robotics: Science and Systems (RSS)*, 2024.

[19] A. Maddukuri, Z. Jiang, L. Y. Chen, S. Nasiriany, Y. Xie, Y. Fang, W. Huang, Z. Wang, Z. Xu, N. Chernyadev, S. Reed, K. Goldberg, A. Mandlekar, L. Fan, and Y. Zhu, "Sim-and-real co-training: A simple recipe for vision-based robotic manipulation," *arXiv preprint arXiv:2503.24361*, 2025.

[20] A. Mandlekar, S. Nasiriany, B. Wen, I. Akinola, Y. Narang, L. Fan, Y. Zhu, and D. Fox, "Mimicgen: A data generation system for scalable robot

learning using human demonstrations," in *7th Annual Conference on Robot Learning*, 2023.

[21] C. Garrett, A. Mandlekar, B. Wen, and D. Fox, "Skillmimicgen: Automated demonstration generation for efficient skill learning and deployment," in *8th Annual Conference on Robot Learning*, 2024.

[22] Z. Jiang, Y. Xie, K. Lin, Z. Xu, W. Wan, A. Mandlekar, L. Fan, and Y. Zhu, "Dexmimicgen: Automated data generation for bimanual dexterous manipulation via imitation learning," 2024.

[23] X. Gu, C. Wen, W. Ye, J. Song, and Y. Gao, "Seer: Language instructed video prediction with latent diffusion models," in *The Twelfth International Conference on Learning Representations*, 2024.

[24] K. Black, M. Nakamoto, P. Atreya, H. R. Walke, C. Finn, A. Kumar, and S. Levine, "Zero-shot robotic manipulation with pre-trained image-editing diffusion models," in *The Twelfth International Conference on Learning Representations*, 2024.

[25] S. Liu, L. Wu, B. Li, H. Tan, H. Chen, Z. Wang, K. Xu, H. Su, and J. Zhu, "Rdt-1b: a diffusion foundation model for bimanual manipulation," 2024.

[26] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3m: A universal visual representation for robot manipulation," in *6th Annual Conference on Robot Learning*, 2022.

[27] S. Karamcheti, S. Nair, A. S. Chen, T. Kollar, C. Finn, D. Sadigh, and P. Liang, "Language-driven representation learning for robotics," in *Robotics: Science and Systems (RSS)*, 2023.

[28] V. Jain, M. Attarian, N. J. Joshi, A. Wahid, D. Driess, Q. Vuong, P. R. Sanketi, P. Sermanet, S. Welker, C. Chan, I. Gilitschenski, Y. Bisk, and D. Dwibedi, "Vid2robot: End-to-end video-conditioned policy learning with cross-attention transformers," in *Robotics: Science and Systems (RSS)*, 2024.

[29] S. Tao, F. Xiang, A. Shukla, Y. Qin, X. Hinrichsen, X. Yuan, C. Bao, X. Lin, Y. Liu, T. kai Chan, Y. Gao, X. Li, T. Mu, N. Xiao, A. Gurha, Z. Huang, R. Calandra, R. Chen, S. Luo, and H. Su, "Maniskill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai," 2024.

[30] J. H. Yang, D. Sadigh, and C. Finn, "Polybot: Training one policy across robots while embracing variability," in *7th Annual Conference on Robot Learning*, 2023.

[31] J. Yang, C. Glossop, A. Bhorkar, D. Shah, Q. Vuong, C. Finn, D. Sadigh, and S. Levine, "Pushing the limits of cross-embodiment learning for manipulation and navigation," 2024.

[32] R. Doshi, H. R. Walke, O. Mees, S. Dasari, and S. Levine, "Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation," in *8th Annual Conference on Robot Learning*, 2024.

[33] H. Geng, F. Wang, S. Wei, Y. Li, B. Wang, B. An, C. T. Cheng, H. Lou, P. Li, Y.-J. Wang, Y. Liang, D. Goetting, C. Xu, H. Chen, Y. Qian, Y. Geng, J. Mao, W. Wan, M. Zhang, J. Lyu, S. Zhao, J. Zhang, J. Zhang, C. Zhao, H. Lu, Y. Ding, R. Gong, Y. Wang, Y. Kuang, R. Wu, B. Jia, C. Sferrazza, H. Dong, S. Huang, K. Sreenath, Y. Wang, J. Malik, and P. Abbeel, "Roboverse: Towards a unified platform, dataset and benchmark for scalable and generalizable robot learning," April 2025. [Online]. Available: https://github.com/RoboVerseOrg/RoboVerse

[34] M. Dalal, M. Liu, W. Talbott, C. Chen, D. Pathak, J. Zhang, and R. Salakhutdinov, "Local policies enable zero-shot long-horizon manipulation," *arXiv preprint arXiv:2410.22332*, 2024.

[35] M. Torne, A. Jain, J. Yuan, V. Macha, L. Ankile, A. Simeonov, P. Agrawal, and A. Gupta, "Robot learning with super-linear scaling," 2024.

[36] P. Katara, Z. Xian, and K. Fragkiadaki, "Gen2sim: Scaling up robot learning in simulation with generative models," 2023.

[37] H. Tan, X. Xu, C. Ying, X. Mao, S. Liu, X. Zhang, H. Su, and J. Zhu, "Manibox: Enhancing spatial grasping generalization via scalable simulation data generation," 2024. [Online]. Available: https://arxiv.org/abs/2411.01850

[38] Y. Wang, Z. Xian, F. Chen, T.-H. Wang, Y. Wang, K. Fragkiadaki, Z. Erickson, D. Held, and C. Gan, "Robogen: Towards unleashing infinite data for automated robot learning via generative simulation," in *Forty-first International Conference on Machine Learning*, 2024.

[39] H. Ha, P. Florence, and S. Song, "Scaling up and distilling down: Language-guided robot skill acquisition," in *7th Annual Conference on Robot Learning*, 2023.

[40] Y. Du, S. Yang, B. Dai, H. Dai, O. Nachum, J. B. Tenenbaum, D. Schuurmans, and P. Abbeel, "Learning universal policies via text-guided video generation," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[41] A. Ajay, S. Han, Y. Du, S. Li, A. Gupta, T. Jaakkola, J. Tenenbaum, L. Kaelbling, A. Srivastava, and P. Agrawal, "Compositional foundation models for hierarchical planning," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 22 304–22 325.

[42] M. Shridhar, Y. L. Lo, and S. James, "Generative image as action models," in *8th Annual Conference on Robot Learning*, 2024.

[43] Y. Du, S. Yang, P. Florence, F. Xia, A. Wahid, brian ichter, P. Sermanet, T. Yu, P. Abbeel, J. B. Tenenbaum, L. P. Kaelbling, A. Zeng, and J. Tompson, "Video language planning," in *The Twelfth International Conference on Learning Representations*, 2024.

[44] S. Yang, J. C. Walker, J. Parker-Holder, Y. Du, J. Bruce, A. Barreto, P. Abbeel, and D. Schuurmans, "Position: Video as the new language for real-world decision making," in *Forty-first International Conference on Machine Learning*, 2024.

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

[46] H. Bharadhwaj, D. Dwibedi, A. Gupta, S. Tulsiani, C. Doersch, T. Xiao, D. Shah, F. Xia, D. Sadigh, and S. Kirmani, "Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation," 2024.

[47] J. Jang, S. Ye, Z. Lin, J. Xiang, J. Bjorck, Y. Fang, F. Hu, S. Huang, K. Kundalia, Y.-C. Lin *et al.*, "Dreamgen: Unlocking generalization in robot learning through video world models," *arXiv preprint arXiv:2505.12705v2*, 2025.

[48] X. Qiu, Y. Wang, J. Cai, Z. Chen, C. Lin, T.-H. Wang, and C. Gan, "Lucibot: Automated robot policy learning from generated videos," 2025. [Online]. Available: https://arxiv.org/abs/2503.09871

[49] S. Patel, S. Mohan, H. Mai, U. Jain, S. Lazebnik, and Y. Li, "Robotic manipulation by imitating generated videos without physical demonstrations," *arXiv preprint arXiv:2507.00990*, 2025.

[50] Z. Chen, Z. Mandi, H. Bharadhwaj, M. Sharma, S. Song, A. Gupta, and V. Kumar, "Semantically controllable augmentations for generalizable robot learning," *The International Journal of Robotics Research*, vol. 0, no. 0, p. 02783649241273686, 0.

[51] T. Yu, T. Xiao, A. Stone, J. Tompson, A. Brohan, S. Wang, J. Singh, C. Tan, D. M, J. Peralta, B. Ichter, K. Hausman, and F. Xia, "Scaling robot learning with semantically imagined experience," in *Robotics: Science and Systems (RSS)*, 2023.

[52] X. Zhang, M. Chang, P. Kumar, and S. Gupta, "Diffusion meets dagger: Supercharging eye-in-hand imitation learning," in *Robotics: Science and Systems (RSS)*, 2024.

[53] G. Authors, "Genesis: A universal and generative physics engine for robotics and beyond," December 2024.

[54] Z. Xian, T. Gervet, Z. Xu, Y.-L. Qiao, T.-H. Wang, and Y. Wang, "Towards generalist robots: A promising paradigm via generative simulation," 2023.

[55] X. Li, K. Hsu, J. Gu, O. Mees, K. Pertsch, H. R. Walke, C. Fu, I. Lunawat, I. Sieh, S. Kirmani, S. Levine, J. Wu, C. Finn, H. Su, Q. Vuong, and T. Xiao, "Evaluating real-world robot manipulation policies in simulation," in *8th Annual Conference on Robot Learning*, 2024.

[56] M. Deitke, R. Liu, M. Wallingford, H. Ngo, O. Michel, A. Kusupati, A. Fan, C. Laforte, V. Voleti, S. Y. Gadre, E. VanderBilt, A. Kembhavi, C. Vondrick, G. Gkioxari, K. Ehsani, L. Schmidt, and A. Farhadi, "Objaverse-XL: A universe of 10m+ 3d objects," in *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[57] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, D. Gordon, Y. Zhu, A. Gupta, and A. Farhadi, "AI2-THOR: an interactive 3d environment for visual AI," *arXiv*, 2017.

[58] J. Collins, S. Goel, A. Luthra, L. Xu, K. Deng, X. Zhang, T. F. Y. Vicente, H. Arora, T. Dideriksen, M. Guillaumin *et al.*, "Abo: Dataset and benchmarks for real-world 3d object understanding," in *CVPR*, 2022.

[59] L. Downs, A. Francis, N. Koenig, B. Kinman, R. Hickman, K. Reymann, T. B. McHugh, and V. Vanhoucke, "Google scanned objects: A high-quality dataset of 3d scanned household items," in *ICRA*, 2022.

[60] M. Khanna*, Y. Mao*, H. Jiang, S. Haresh, B. Shacklett, D. Batra, A. Clegg, E. Undersander, A. X. Chang, and M. Savva, "Habitat Synthetic Scenes Dataset (HSSD-200): An Analysis of 3D Scene Scale and Realism Tradeoffs for ObjectGoal Navigation," *arXiv preprint*, 2023.