Interactive World Simulator for Robot Policy Training and Evaluation

Yixuan Wang[1] Rhythm Syed[1] Fangyu Wu[1] Mengchao Zhang[2] Aykut Onol[2] Jose Barreiros[2]
Hooshang Nayyeri[3] Tony Dear[1] Huan Zhang[4] Yunzhu Li[1]
[1]Columbia University [2]Toyota Research Institute [3]Amazon [4]University of Illinois Urbana-Champaign
https://www.yixuanwang.me/interactive_world_sim/

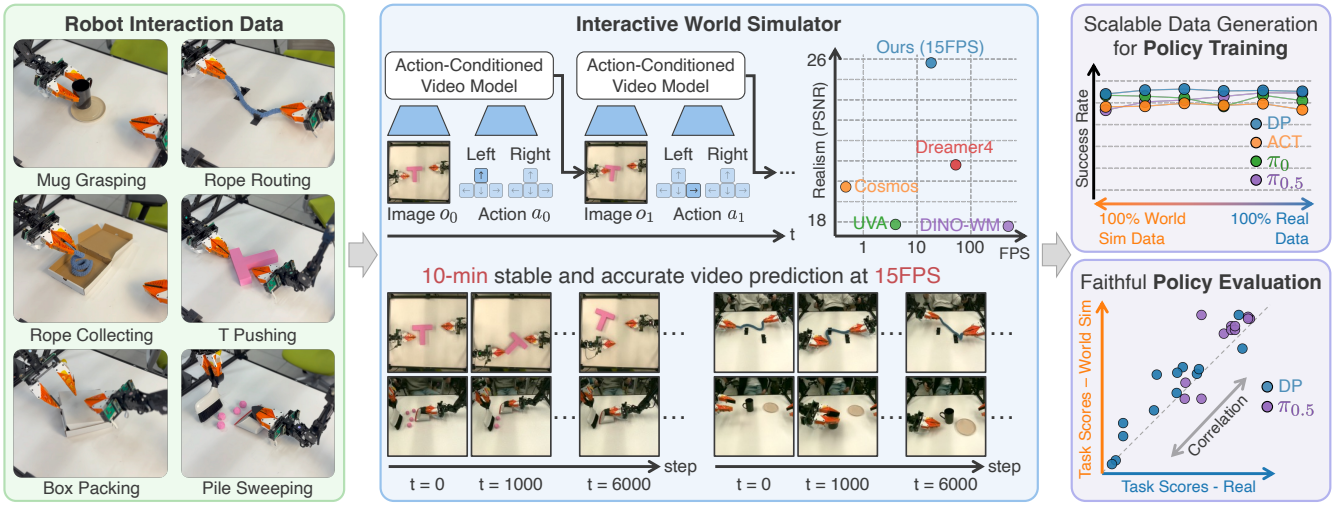

Fig. 1: **Overview of the Interactive World Simulator.** Given real-world robot interaction datasets (left), we train action-conditioned video models that capture complex physical dynamics and support stable long-horizon interactions (middle). The resulting world models achieve stable video prediction for over **10 minutes at 15 FPS on a single RTX 4090 GPU**, outperforming prior world models in long-horizon realism. The resulting Interactive World Simulator enables **scalable data generation for imitation policy training** without requiring additional robot interaction and supports **faithful and reproducible policy evaluation**, exhibiting a strong correlation between policy performance in simulation and in the real world (right). Additional examples are available on the project website.

***Abstract*—Action-conditioned video prediction models (often referred to as world models) have shown strong potential for robotics applications, but existing approaches are often slow and struggle to capture physically consistent interactions over long horizons, limiting their usefulness for scalable robot policy training and evaluation. We present Interactive World Simulator, a framework for building interactive world models from a moderate-sized robot interaction dataset. Our approach leverages consistency models for both image decoding and latent-space dynamics prediction, enabling fast and stable simulation of physical interactions. In our experiments, the learned world models produce interaction-consistent pixel-level predictions and support stable long-horizon interactions for more than *10 minutes at 15 FPS on a single RTX 4090 GPU*. Our framework enables scalable demonstration collection solely within the world models to train state-of-the-art imitation policies. Through extensive real-world evaluation across diverse tasks involving rigid objects, deformable objects, object piles, and their interactions, we find that policies trained on world-model-generated data perform comparably to those trained on the same amount of real-world data. Additionally, we evaluate policies both within the world models and in the real world across diverse tasks, and observe a strong correlation between simulated and real-world performance. Together, these results establish the Interactive World Simulator as a stable and physically consistent surrogate for scalable robotic data generation and faithful, reproducible policy evaluation.**

I. Introduction

Video prediction models have shown promising results for robotic manipulation, including planning [1], control [2, 3], policy steering [4], and policy evaluation [5]. However, existing action-conditioned video prediction models are either computationally expensive, due to heavy neural networks and multi-step diffusion processes [6–8], or unstable over long-horizon rollouts due to accumulated prediction errors [2, 3]. As a result, faithful long-horizon action-conditioned video prediction of complex physical interactions remains challenging for state-of-the-art models, limiting their applicability to scalable policy training data generation and reproducible policy evaluation.

In this work, we introduce *Interactive World Simulator* (Figure 1), an action-conditioned video prediction model that

supports stable, interactive rollouts for more than **10 minutes at 15 FPS on a single RTX 4090 GPU**. To achieve efficient long-horizon prediction, we first encode high-dimensional images into compact 2D latent representations and perform future prediction entirely in latent space. In the first stage, we train an autoencoder using a CNN encoder [9] and a consistency-model decoder [10] to enable efficient and high-fidelity reconstruction. In the second stage, we freeze the autoencoder and train an action-conditioned dynamics model using next-frame supervision in latent space. We model latent dynamics using consistency models, which are computationally efficient and capable of representing the multimodal distribution of possible future outcomes arising from robotic interaction. During inference, we autoregressively shift a fixed-length context window to generate long-horizon video predictions conditioned on robot actions, as shown in Figure 2.

Our Interactive World Simulator enables two important robotic applications: **1) Scalable high-quality data generation for imitation policy training.** Imitation learning has demonstrated strong performance in robotic manipulation [11–14], but existing approaches typically require large amounts of real-robot data, which are expensive to collect and difficult to scale. Our world simulator provides a realistic interactive environment in which users can collect demonstration data directly through interaction, enabling large-scale data collection without access to physical robots and substantially reducing data collection cost. **2) Reproducible policy evaluation within the Interactive World Simulator.** Policy evaluation is a longstanding challenge in robotics and is critical for algorithm iteration and checkpoint selection. As noted in [15], real-world evaluation is time-consuming and difficult to conduct in a controlled, apples-to-apples manner, which slows policy development and complicates fair comparison between methods. In contrast, policy evaluation within our world simulator is scalable and reproducible, enabling efficient and consistent comparison across policies. Because our model is trained on real-world robot interaction data, it exhibits reduced domain gap and dynamics mismatch compared to conventional simulators.

We conduct comprehensive experiments comparing our action-conditioned video prediction model with state-of-the-art baselines across a diverse set of tasks involving rigid objects, deformable objects, articulated objects, object piles, and their interactions, as shown in Figure 1. Our model outperforms prior methods, including Cosmos [6], UVA [8], DINO-WM [3], and Dreamer4 [16], in terms of long-horizon prediction realism, while maintaining **interactive performance at 15 FPS on a single RTX 4090 GPU**.

To study whether our world simulator can generate data of comparable quality to real-world demonstrations, we collect the same amount of expert data within the simulator under identical initial configurations and data collection protocols. We then train imitation policies, including Diffusion Policy (DP), Action Chunking Transformer (ACT), π_0, and $\pi_{0.5}$ [11–14], using different mixtures of real-world and simulator-generated data. Across all mixture ratios, ranging from 100% world-simulator data to 100% real-world data, we observe comparable policy performance, indicating that simulator-generated data are of similar quality to real-world data. Finally, we evaluate policies both in the real world and in the world simulator across multiple tasks and training checkpoints, and observe a strong correlation between simulator and real-world performance, demonstrating that our world simulator can faithfully reflect relative policy performance.

In summary, our main contributions are threefold: 1) we introduce the *Interactive World Simulator*, an interactive action-conditioned video prediction model that supports stable long-horizon rollouts for over 10 minutes at 15 FPS across complicated physical interactions involving rigid objects, deformable objects, object piles, and multi-object interactions; 2) using this simulator, we enable scalable, high-quality data collection for imitation learning without requiring access to physical robots; and 3) we demonstrate through extensive experiments that policy performance in the world simulator strongly correlates with real-world performance, enabling reproducible and scalable policy evaluation.

II. Method

In this section, we first describe our problem formulation in Section II-A. We then elaborate on model training and inference in Section II-B. In Section II-C and Section II-D, we discuss how to use this model for scalable data generation and reproducible policy evaluation.

A. Problem Formulation

We assume that the robotic interaction dataset $\mathcal{D}$ consists of multiple episodes, each of which has the form $\mathcal{E} = \{(o_0, a_0), (o_1, a_1), ..., (o_T, a_T)\}$, where $o_t \in \mathbb{R}^{3\times H\times W}$ is an RGB image at time t and a_t is the action at time t. Our goal is to build an action-conditioned video prediction model $\hat{o}_t = f(o_{t-N:t-1}, a_{t-N:t-1})$ that takes in N history observations and predicts the next RGB frame $\hat{o}_t$ by minimizing $||\hat{o}_t - o_t||_2$, the difference from the ground truth observation.

B. Interactive World Simulator

As shown in Figure 2, we train our action-conditioned video prediction model in two stages. In the first stage, we train an autoencoder that can encode an image o into a latent representation z and decode the latent representation to reconstruct the input image. In the second stage, we freeze the encoder and decoder and train the action-conditioned dynamics model using next-frame supervision in latent space. We elaborate on the different stages in the following section.

Consistency models are widely used for image generation due to their efficiency and quality [10, 17, 18]. Therefore, we instantiate the decoder as a consistency model for high-quality image generation. Due to instability of 1-step consistency model training, we are inspired by Consistency Trajectory Model (CTM) for stable consistency model training [18].

After stage 1, we freeze the autoencoder parameters (ϕ, θ) and encode each frame o_t into a latent z_t. Our goal in stage 2 is to learn an action-conditioned latent dynamics model F_ψ

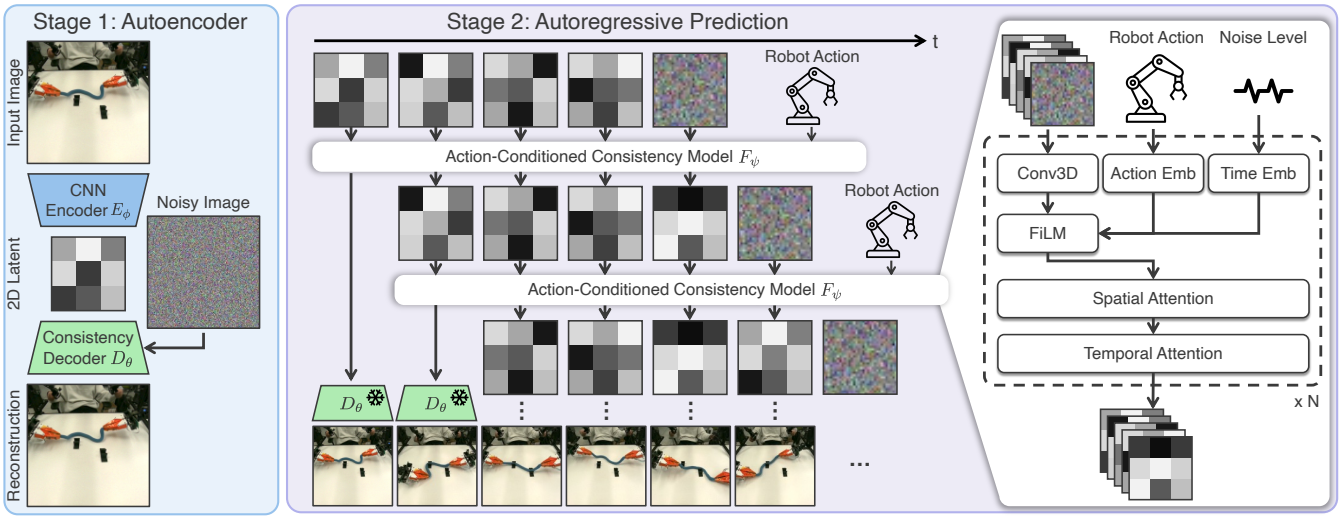

Fig. 2: **Method Overview.** Model training proceeds in two stages. In Stage 1, we train an autoencoder that maps RGB observations into compact 2D latent representations using a CNN encoder E_ϕ and a consistency-model decoder D_θ, enabling efficient and high-fidelity reconstruction. In Stage 2, we freeze the autoencoder and train an action-conditioned consistency model F_ψ in latent space to model future visual dynamics using next-frame supervision. We choose F_ψ as a consistency model because it efficiently represents multimodal distributions over possible future outcomes. The dynamics model learns to denoise noisy latent states conditioned on robot actions, past latents, and noise levels, and is instantiated as a stack of 3D convolutional blocks with FiLM modulation and spatiotemporal attention. During inference, F_ψ autoregressively predicts future latents, which are decoded by D_θ to generate long-horizon video predictions.

that predicts future latent frames given a context window of past latents $z_{t-N:t-1}$ and actions $a_{t-N:t-1}$. Since consistency models can naturally model the multimodal distribution of potential futures while being efficient, we also model F_ψ as a consistency model.

During inference time, we predict future frames autoregressively. Given initial image o_0, we first obtain 2D latent $z_0 = E_\phi(o_0)$. As illustrated by Figure 2, we attach a noisy latent to history latents. Alongside action information, we denoise the last-frame noisy latent into a clean one (i.e., z_1). Then we attach the latest predicted latent to existing context to get the new history context (e.g., $z_{0:1}$). By repeating this process, we are able to autoregressively predict latents. Note that we discard the old latent once the context length is larger than a threshold so that the computation cost will not increase as horizon increases. Finally, the decoder D_θ takes in newly predicted latents z and renders new images $\hat{o}$.

C. Data Generation for Policy Training

Our Interactive World Simulator serves as a scalable surrogate for expert demonstration data collection, enabling the generation of high-quality synthetic demonstrations for imitation policy training without requiring access to physical robots.

D. Policy Evaluation

Evaluating policies fairly is another longstanding challenge in robotics. Performing real-world evaluation is time-consuming and difficult to scale, as each trial requires manual resets and careful control of experimental conditions such as the initial object configuration. Our Interactive World Simulator enables scalable policy evaluation by allowing policies to interact directly with the simulated environment: the world model takes policy actions and predicts future frames, while the policy consumes the predicted frames to generate subsequent actions, mirroring real-world interaction. Compared to physical experiments, interacting within the world simulator enables controllable initial configurations, faster experiment resets, and safer large-scale evaluation. In the experiments section, we study whether policy performance in the world simulator faithfully predicts performance in the real world.

III. Experiment

In this section, we aim to study three questions: 1) How does our Interactive World Simulator perform compared to state-of-the-art world models in terms of realism, speed, and robustness? (Section III-B) 2) Can our Interactive World Simulator generate data with quality similar to real data for imitation learning? (Section III-C) 3) Can our Interactive World Simulator's policy evaluation faithfully represent real-world policy performance? (Section III-D)

A. World Model Instantiation

We set up one simulation task within MuJoCo [19] and six real-world tasks, including `Mug Grasping`, `Rope Routing`, `Rope Collecting`, `T Pushing`, `Box Packing`, and `Pile Sweeping`, which involve rigid objects, deformable objects, object piles, articulated objects, and the interactions among them. We performed all tasks using the ALOHA Bimanual Robot [12]. In simulation, we instantiate the T pushing task similarly to its real-world counterpart, as shown in Figure 1. We use a scripted policy to generate 10,000 episodes of random interactions to ensure sufficient data coverage. In the real world, we collect about 600 episodes of play data for each task, with each episode consisting of 200

steps. Real-world data collection for each task takes around six hours for a single person, which is a manageable scale for most research labs.

B. Video Baseline Comparisons

To evaluate the video prediction performance of the proposed Interactive World Simulator against prior action-conditioned world models, we conduct long-horizon action-conditioned video prediction comparisons across all six real tasks and one simulation task listed in Section III-A. We compare against representative state-of-the-art baselines, including Cosmos [6], UVA [8], Dreamer4 [16], and DINO-WM [3], under identical initial conditions and rollout horizons (192 steps or 19.2 seconds). Prediction quality is assessed using standard reconstruction, perceptual, and temporal metrics such as MSE, PSNR, and FVD, computed between predicted and ground-truth video rollouts.

C. Data Generation for Policy Training

To evaluate whether our world simulator can serve as a scalable data generator, we investigate whether synthetic demonstrations provide similar quality to real-world robot data for training imitation policies. We consider five manipulation tasks: T pushing in MuJoCo, T pushing in the real world, and three additional real-world tasks, pile sweeping, mug grasping, and rope routing. We benchmark four imitation learning policies: DP, ACT, π_0, and $\pi_{0.5}$. For each task, we use exactly 100 demonstration episodes to train imitation policies, regardless of the data source.

Our analysis confirms that policies trained on 100% world simulator data achieve performance levels comparable to those trained on 100% real world expert data. For DP, the average task scores remain extremely stable, with a score of 87.9% using 100% simulator data compared to 90.3% using 100% real data. Similarly, ACT demonstrates strong parity, achieving 76.2% using simulator data versus 73.6% using real data. While $\pi_{0.5}$ shows a general trend of improvement with increasing real data, rising from 73.1% to 88.8%, it still achieves robust performance among the variety of tasks and initial configurations evaluated across simulation and real environments. Notably, policies trained on 100% world simulator data perform comparably to those trained on an equivalent volume of real-robot expert data. This suggests that our simulator can generate data with quality similar to that of real-world demonstrations.

D. Sim-to-Real Correlation for Faithful Policy Evaluation

We investigate whether our world simulator can serve as a proxy for real-world policy evaluation and faithfully represent real-world performance. For this to hold, policy performance in the world simulator must be positively correlated with performance in the real world under identical initial conditions.

We train four imitation policies, DP, ACT, π_0, and $\pi_{0.5}$, on real-world data and evaluate the final and intermediate checkpoints of the policies on four manipulation tasks: T pushing, rope routing, mug grasping, and pile sweeping. For each task, we sample 20 initial configurations from the world simulator's training distribution and deploy each policy in both the world simulator and the real world.

Across all four tasks, we observe strong positive correlations between simulator and real-world performance. For tasks other than T pushing, the fitted lines exhibit a slight positive bias, indicating that policies tend to achieve slightly higher scores in the world simulator than in the real world. Despite the sim-to-real gap, this bias does not diminish the simulator's utility: if one policy substantially outperforms another in the world simulator, our results suggest that it is likely to do so in the real world as well. This property is useful in practice: one can rely on the world simulator to select high-quality candidate policies without the need for real-world experiments, thus reducing development cost and iteration cycles.

IV. Conclusion

While action-conditioned video prediction shows great potential for robotics, existing approaches are often either too slow for interactive use or brittle under long-horizon inference. In this work, we introduced the Interactive World Simulator, an action-conditioned video world model that supports long-horizon, stable, and physically consistent robot interactions. The simulator enables efficient video prediction for over 10 minutes of continuous interaction at 15 FPS while maintaining high-quality pixel-level predictions. Our approach addresses key limitations of prior world models for robotics, improving both stability and computational efficiency.

Our experiments demonstrate the dual utility of the simulator as a scalable engine for both policy training and evaluation. We show that imitation policies trained exclusively on simulator-generated data perform comparably to those trained on real-world expert demonstrations. Furthermore, we observe strong correlations between policy performance in the simulator and in the real world across several tasks, validating the Interactive World Simulator as a reliable surrogate for policy training and evaluation.

Looking forward, we plan to extend this framework to more diverse environments and increasingly complex manipulation tasks, further unlocking the potential of large-scale robotic data. An important direction for future work is to study how the performance of world models scales with increasing amounts of interaction data and computational resources. Understanding these scaling behaviors could guide the design of larger and more capable world models for robotics.

References

[1] Boyuan Chen, Tianyuan Zhang, Haoran Geng, Kiwhan Song, Caiyi Zhang, Peihao Li, William T Freeman, Jitendra Malik, Pieter Abbeel, Russ Tedrake, et al. Large video planner enables generalizable robot control. *arXiv preprint arXiv:2512.15840*, 2025.

[2] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

[3] Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024.

[4] Yilin Wu, Ran Tian, Gokul Swamy, and Andrea Bajcsy. From foresight to forethought: Vlm-in-the-loop policy steering via latent alignment. *arXiv preprint arXiv:2502.01828*, 2025.

[5] Gemini Robotics Team, Krzysztof Choromanski, Coline Devin, Yilun Du, Debidatta Dwibedi, Ruiqi Gao, Abhishek Jindal, Thomas Kipf, Sean Kirmani, Isabel Leal, et al. Evaluating gemini robotics policies in a veo world simulator. *arXiv preprint arXiv:2512.10675*, 2025.

[6] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

[7] Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-world: A controllable generative world model for robot manipulation. *arXiv preprint arXiv:2510.10125*, 2025.

[8] Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. *arXiv preprint arXiv:2503.00200*, 2025.

[9] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002.

[10] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, pages 32211–32252. PMLR, 2023.

[11] 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.

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

[13] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. pi0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

[14] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. pi0.5: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

[15] Jose Barreiros, Andrew Beaulieu, Aditya Bhat, Rick Cory, Eric Cousineau, Hongkai Dai, Ching-Hsin Fang, Kunimatsu Hashimoto, Muhammad Zubair Irshad, Masha Itkina, et al. A careful examination of large behavior models for multitask dexterous manipulation. *arXiv preprint arXiv:2507.05331*, 2025.

[16] Danijar Hafner, Wilson Yan, and Timothy Lillicrap. Training agents inside of scalable world models. *arXiv preprint arXiv:2509.24527*, 2025.

[17] Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. *arXiv preprint arXiv:2310.14189*, 2023.

[18] Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. *arXiv preprint arXiv:2310.02279*, 2023.

[19] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE, 2012.

[20] Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos J Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. *Advances in Neural Information Processing Systems*, 37:58757–58791, 2024.

[21] Philip J Ball, Jakob Bauer, Frank Belletti, B Brownfield, A Ephrat, S Fruchter, A Gupta, K Holsheimer, A Holynski, J Hron, et al. Genie 3: A new frontier for world models. *Google DeepMind Blog*, pages 253–279, 2025.

[22] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024.

[23] Junbang Liang, Ruoshi Liu, Ege Ozguroglu, Sruthi Sudhakar, Achal Dave, Pavel Tokmakov, Shuran Song, and Carl Vondrick. Dreamitate: Real-world visuomotor policy learning via video generation. In *CoRL 2024*, 2024. URL https://dreamitate.cs.columbia.edu.

[24] Junbang Liang, Pavel Tokmakov, Ruoshi Liu, Sruthi Sudhakar, Paarth Shah, Rares Ambrus, and Carl Vondrick. Video generators are robot policies. *arXiv 2025*, 2025. URL https://arxiv.org/abs/2508.00795.

[25] Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, et al. Cosmos policy: Fine-tuning video models for visuomotor control and planning. *arXiv preprint arXiv:2601.16163*, 2026.

[26] Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang,

Kaushil Kundalia, Yen-Chen Lin, et al. Dreamgen: Unlocking generalization in robot learning through video world models. *arXiv preprint arXiv:2505.12705*, 2025.

[27] Zhiting Mei, Tenny Yin, Ola Shorinwa, Apurva Badithela, Zhonghe Zheng, Joseph Bruno, Madison Bland, Lihan Zha, Asher Hancock, Jaime Fernández Fisac, et al. Video generation models in robotics-applications, research challenges, future directions. *arXiv preprint arXiv:2601.07823*, 2026.

[28] Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025.

[29] Alejandro Escontrela, Ademi Adeniji, Wilson Yan, Ajay Jain, Xue Bin Peng, Ken Goldberg, Youngwoon Lee, Danijar Hafner, and Pieter Abbeel. Video prediction models as rewards for reinforcement learning. *Advances in Neural Information Processing Systems*, 36:68760–68783, 2023.

[30] Katerina Fragkiadaki, Pulkit Agrawal, Sergey Levine, and Jitendra Malik. Learning visual predictive models of physics for playing billiards. *arXiv preprint arXiv:1511.07404*, 2015.

[31] Pulkit Agrawal, Ashvin V Nair, Pieter Abbeel, Jitendra Malik, and Sergey Levine. Learning to poke by poking: Experiential learning of intuitive physics. *Advances in neural information processing systems*, 29, 2016.

[32] Himanshu Gaurav Singh, Pieter Abbeel, Jitendra Malik, and Antonio Loquercio. Deep sensorimotor control by imitating predictive models of human motion. *arXiv preprint arXiv:2508.18691*, 2025.

[33] Arthur Allshire, Hongsuk Choi, Junyi Zhang, David McAllister, Anthony Zhang, Chung Min Kim, Trevor Darrell, Pieter Abbeel, Jitendra Malik, and Angjoo Kanazawa. Visual imitation enables contextual humanoid control. *arXiv preprint arXiv:2505.03729*, 2025.

[34] Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Advances in neural information processing systems*, 36:9156–9172, 2023.

[35] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pretraining for visual robot manipulation. *arXiv preprint arXiv:2312.13139*, 2023.

[36] Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, Hanbo Zhang, and Minzhao Zhu. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*, 2024.

[37] Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, Lars Liden, Kimin Lee, Jianfeng Gao, Luke Zettlemoyer, Dieter Fox, and Minjoon Seo. Latent action pretraining from videos, 2024. URL https://arxiv.org/abs/2410.11758.

[38] Chuan Wen, Xingyu Lin, John So, Kai Chen, Qi Dou, Yang Gao, and Pieter Abbeel. Any-point trajectory modeling for policy learning. *arXiv preprint arXiv:2401.00025*, 2023.

[39] Kanchana Ranasinghe, Honglu Zhou, Yu Fang, Luyu Yang, Le Xue, Ran Xu, Caiming Xiong, Silvio Savarese, Michael S Ryoo, and Juan Carlos Niebles. Future optical flow prediction improves robot control & video generation. *arXiv preprint arXiv:2601.10781*, 2026.

[40] Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Malik, Seonghyeon Ye, Sihyun Yu, Wei-Cheng Tseng, Yuzhu Dong, Kaichun Mo, Chen-Hsuan Lin, et al. Dreamdojo: A generalist robot world model from large-scale human videos. *arXiv preprint arXiv:2602.06949*, 2026.

[41] Moritz Reuss, Ömer Erdinç Yağmurlu, Fabian Wenzel, and Rudolf Lioutikov. Multimodal diffusion transformer: Learning versatile behavior from multimodal goals. *arXiv preprint arXiv:2407.05996*, 2024.

[42] Yang Tian, Sizhe Yang, Jia Zeng, Ping Wang, Dahua Lin, Hao Dong, and Jiangmiao Pang. Predictive inverse dynamics models are scalable learners for robotic manipulation. *arXiv preprint arXiv:2412.15109*, 2024.

[43] Ruijie Zheng, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loic Magne, et al. Flare: Robot learning with implicit world modeling. *arXiv preprint arXiv:2505.15659*, 2025.

[44] Yanjiang Guo, Yucheng Hu, Jianke Zhang, Yen-Jen Wang, Xiaoyu Chen, Chaochao Lu, and Jianyu Chen. Prediction with action: Visual policy learning via joint denoising process. *Advances in Neural Information Processing Systems*, 37:112386–112410, 2024.

[45] Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. *arXiv preprint arXiv:2412.14803*, 2024.

[46] Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025.

[47] Jiayi Chen, Wenxuan Song, Pengxiang Ding, Ziyang Zhou, Han Zhao, Feilong Tang, Donglin Wang, and Haoang Li. Unified diffusion vla: Vision-language-action model via joint discrete denoising diffusion process. *arXiv preprint arXiv:2511.01718*, 2025.

[48] Sandeep Routray, Hengkai Pan, Unnat Jain, Shikhar Bahl, and Deepak Pathak. Vipra: Video prediction for robot actions. *arXiv preprint arXiv:2511.07732*, 2025.

[49] Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, et al. Causal world modeling for robot control.

arXiv preprint arXiv:2601.21998, 2026.

[50] Hongyu Li, Lingfeng Sun, Yafei Hu, Duy Ta, Jennifer Barry, George Konidaris, and Jiahui Fu. Novaflow: Zero-shot manipulation via actionable flow from generated videos. *arXiv preprint arXiv:2510.08568*, 2025.

[51] Shivansh Patel, Shraddhaa Mohan, Hanlin Mai, Unnat Jain, Svetlana Lazebnik, and Yunzhu Li. Robotic manipulation by imitating generated videos without physical demonstrations. *arXiv preprint arXiv:2507.00990*, 2025.

[52] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[53] Aäron van den Oord and Elias Roman. State-of-the-art video and image generation with Veo 2 and Imagen 3. Google DeepMind Blog, 2024. https://deepmind.google/blog/.

[54] Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024.

[55] Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. *arXiv preprint arXiv:2502.06764*, 2025.

[56] Jonas Pai, Liam Achenbach, Victoriano Montesinos, Benedek Forrai, Oier Mees, and Elvis Nava. mimic-video: Video-action models for generalizable robot control beyond vlas. *arXiv preprint arXiv:2512.15692*, 2025.

[57] Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026.

[58] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

[59] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.

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

[61] Jason Lee, Jiafei Duan, Haoquan Fang, Yuquan Deng, Shuo Liu, Boyang Li, Bohan Fang, Jieyu Zhang, Yi Ru Wang, Sangho Lee, et al. Molmoact: Action reasoning models that can reason in space. *arXiv preprint arXiv:2508.07917*, 2025.

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

[63] Gemini Robotics Team, S Abeyruwan, J Ainslie, JB Alayrac, MG Arenas, T Armstrong, A Balakrishna, R Baruch, M Bauza, M Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *URL https://arxiv.org/abs/2503.20020*, 2025.

[64] Kaifeng Zhang, Shuo Sha, Hanxiao Jiang, Matthew Loper, Hyunjong Song, Guangyan Cai, Zhuo Xu, Xiaochen Hu, Changxi Zheng, and Yunzhu Li. Real-to-sim robot policy evaluation with gaussian splatting simulation of soft-body interactions. *arXiv preprint arXiv:2511.04665*, 2025.

[65] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

[66] Apurva Badithela, David Snyder, Lihan Zha, Joseph Mikhail, Matthew O'Kelly, Anushri Dixit, and Anirudha Majumdar. Reliable and scalable robot policy evaluation with imperfect simulators. *arXiv preprint arXiv:2510.04354*, 2025.

[67] Pranav Atreya, Karl Pertsch, Tony Lee, Moo Jin Kim, Arhan Jain, Artur Kuramshin, Clemens Eppner, Cyrus Neary, Edward Hu, Fabio Ramos, et al. Roboarena: Distributed real-world evaluation of generalist robot policies. *arXiv preprint arXiv:2506.18123*, 2025.

[68] Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, et al. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*, 2024.

[69] Zhiyuan Zhou, Pranav Atreya, You Liang Tan, Karl Pertsch, and Sergey Levine. Autoeval: Autonomous evaluation of generalist robot manipulation policies in the real world. *arXiv preprint arXiv:2503.24278*, 2025.

[70] Xuning Yang, Clemens Eppner, Jonathan Tremblay, Dieter Fox, Stan Birchfield, and Fabio Ramos. Robot policy evaluation for sim-to-real transfer: A benchmarking perspective. *arXiv preprint arXiv:2508.11117*, 2025.

[71] Yi Ru Wang, Carter Ung, Grant Tannert, Jiafei Duan, Josephine Li, Amy Le, Rishabh Oswal, Markus Grotz, Wilbert Pumacay, Yuquan Deng, et al. Roboeval: Where robotic manipulation meets structured and scalable evaluation. *arXiv preprint arXiv:2507.00435*, 2025.

[72] Hadas Kress-Gazit, Kunimatsu Hashimoto, Naveen Kuppuswamy, Paarth Shah, Phoebe Horgan, Gordon Richardson, Siyuan Feng, and Benjamin Burchfiel. Robot learning as an empirical science: Best practices for policy evaluation. *arXiv preprint arXiv:2409.09491*, 2024.

[73] Julian Quevedo, Ansh Kumar Sharma, Yixiang Sun, Varad Suryavanshi, Percy Liang, and Sherry Yang. Worldgym: World model as an environment for policy evaluation. *arXiv preprint arXiv:2506.00613*, 2025.

[74] Lukas Meyer, Floris Erich, Yusuke Yoshiyasu, Marc Stamminger, Noriaki Ando, and Yukiyasu Domae. Pegasus: Physically enhanced gaussian splatting simulation system for 6dof object pose dataset generation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10710–10715. IEEE, 2024.

[75] 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.

[76] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11097–11107, 2020.

[77] NVIDIA. NVIDIA Isaac Sim: Robotics Simulation and Synthetic Data Generation, 2024.

[78] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[79] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[80] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *arXiv preprint arXiv:2406.11838*, 2024.

Appendix A
Related Works

A. Video Prediction Model for Robotic Manipulation

Video prediction models, or world models, have shown significant potential in robotics [6, 20–49], serving critical roles in zero-shot planning [1, 50], policy steering [4], evaluation [5], and direct control. Frameworks like NovaFlow [50] and RIGVid [51] demonstrate the power of repurposing large-scale video generation to derive actionable 3D object flow and trajectories for manipulation without real-world demonstrations. Recent advancements such as the Large Video Planner [1] further establish video as a primary modality for robot foundation models, enabling generative planning in pixel space. These models are also utilized for run-time policy steering. FOREWARN [4] uses an action-conditioned world model alongside a Vision-Language model (VLM) to select optimal plans. Additionally, Ctrl-World [7] and VEO [5] leverage large-capacity video models for evaluation such as Stable Video Diffusion (SVD) [52] and VEO2 [53] respectively, providing high-fidelity simulation of nominal and out-of-distribution scenarios to assess safety and performance.

However, existing video prediction models face significant practical limitations. Many state-of-the-art architectures, including closed-source models such as Sora [22] and open-source frameworks such as Diffusion Forcing [54] and the Diffusion Forcing Transformer (DFoT) [55], are often not explicitly conditioned on robot actions, which is essential for grounding models in physical interaction. Furthermore, many high-capacity diffusion-based models [6–8, 56, 57] are not efficient enough for real-time interaction, often requiring enterprise-level GPU clusters that are inaccessible especially to many academic labs. Finally, some of existing models [2, 3, 58] lack the robustness required for stable, long-horizon prediction. In contrast to the aforementioned methods, our Interactive World Simulator produces physically accurate pixel-level predictions, supports stable interactions for over 10 minutes at 15 FPS on a single consumer RTX 4090 GPU, making it highly accessible.

B. Imitation Policy Training

Imitation Learning (IL) has demonstrated remarkable results across a variety of robotic tasks through architectures such as Diffusion Policy [11], Action Chunking Transformer (ACT) [12], π_0 [13] and $\pi_{0.5}$ [14] and many others [15, 59–64]. Generalist models like Octo [65] have pushed the boundaries of what IL can achieve in unstructured environments. Despite this progress, these methods remain dependent on high-quality real-robot expert data, which are scarce, expensive to collect, and difficult to scale without constant access to physical hardware. Our framework addresses this bottleneck by providing a system to build interactive world models using a play-data interaction dataset. We demonstrate that imitation policies trained on data generated using our simulator perform comparably to those trained on real-world data, enabling researchers and enthusiasts alike to iterate and scale their data collection strategies without solely relying on physical robots.

C. Imitation Policy Evaluation

Reliable policy evaluation is a cornerstone of robotics research [66–73], with real-world deployment serving as the gold standard for assessing performance. However, real-world evaluation is costly, time-consuming, and difficult to scale, making it impractical for frequent algorithm iteration, checkpoint selection, and fair comparison across methods. As a result, there has been growing interest in developing scalable evaluation frameworks whose outcomes correlate well with real-world policy performance. Prior work has explored both 3D- and 2D-based approaches to address this challenge. 3D-based evaluation methods, such as Physically Embodied Gaussian Splatting (PEGASUS) [74], integrate geometry and physics to create digital twins for real-time interaction. Similarly, Real-to-Sim Eval [64] focuses on structured representations for soft-body interactions, but often carries environmental assumptions such as tabletop configurations. Although 2D-based evaluation systems like VEO [5] offer a more flexible approach, they are often closed-source and not readily available for broad academic use. The research community has also introduced various simulation benchmarks to enable systematic study. Traditional physics engines like MuJoCo [19], robosuite [75], and others [76, 77] provide foundational environments for large-scale training and evaluation. We introduce a framework that requires only paired 2D RGB images and actions, making it highly transferable and scalable across diverse manipulation tasks. By ensuring a strong correlation between performance in our simulator and the real world, we provide an open-source, accessible solution for all labs to conduct reproducible and accurate policy evaluation.

Appendix B
Method

In this section, we first describe our problem formulation in Section B-A. We then elaborate on model training and inference in Section B-B. In Section B-C and Section B-D, we discuss how to use this model for scalable data generation and reproducible policy evaluation.

A. Problem Formulation

We assume that the robotic interaction dataset $\mathcal{D}$ consists of multiple episodes, each of which has the form $\mathcal{E} = \{(o_0, a_0), (o_1, a_1), ..., (o_T, a_T)\}$, where $o_t \in \mathbb{R}^{3\times H\times W}$ is an RGB image at time t and a_t is the action at time t. Our goal is to build an action-conditioned video prediction model $\hat{o}_t = f(o_{t-N:t-1}, a_{t-N:t-1})$ that takes in N history observations and predicts the next RGB frame $\hat{o}_t$ by minimizing $||\hat{o}_t - o_t||_2$, the difference from the ground truth observation.

B. Interactive World Simulator

As shown in Figure 2, we train our action-conditioned video prediction model in two stages. In the first stage, we train an autoencoder that can encode an image o into a latent representation z and decode the latent representation to reconstruct the input image. In the second stage, we freeze the encoder and decoder and train the action-conditioned

dynamics model using next-frame supervision in latent space. We elaborate on the different stages in the following section.

1) Stage 1: Autoencoder Training: Consistency models are widely used for image generation due to their efficiency and quality [10, 17, 18]. Therefore, we instantiate the decoder as a consistency model for high-quality image generation. Due to instability of 1-step consistency model training, we are inspired by Consistency Trajectory Model (CTM) for stable consistency model training [18].

To train the autoencoder and find its parameters (ϕ, θ), for each training image o, we first obtain the 2D latent representations $z = E_\phi(o) \in \mathbb{R}^{C \times H' \times W'}$, where C is the latent channel. We then apply the same noise to o at two different sampled scales $\sigma_t > \sigma_s \geq 0$.

$$x_{\sigma_t} = \mathcal{N}(o; \sigma_t), \qquad x_{\sigma_s} = \mathcal{N}(o; \sigma_s), \tag{1}$$

where $\mathcal{N}(\cdot; \sigma)$ denotes the forward noising operator at noise scale σ. The decoder is trained to map the higher-noise input to the lower-noise target conditioned on z:

$$\hat{x}_{\sigma_s} = D_\theta(x_{\sigma_t}; \sigma_t, \sigma_s, z)\,, \tag{2}$$

by minimizing a weighted regression loss

$$\mathcal{L}_{\text{AE}} = \mathbb{E}_{o, \sigma_t > \sigma_s}\Big[w(\sigma_t)\ \|\hat{x}_{\sigma_s} - x_{\sigma_s}\|_2^2\Big]\,, \tag{3}$$

where $w(\sigma)$ is a noise-dependent weight to balance learning across noise scales. Thus, we obtain an image-to-latent encoder $E_\phi(o)$ and a conditional generative decoder $D_\theta(\cdot;\, z)$ that can reconstruct high-fidelity images from noisy inputs with a small number of denoising steps.

2) Stage 2: Dynamics Training: After stage 1, we freeze the autoencoder parameters (ϕ, θ) and encode each frame o_t into a latent z_t. Our goal in stage 2 is to learn an action-conditioned latent dynamics model F_ψ that predicts future latent frames given a context window of past latents $z_{t-N:t-1}$ and actions $a_{t-N:t-1}$. Since consistency models can naturally model the multimodal distribution of potential futures while being efficient, we also model F_ψ as a consistency model.

Concretely, we treat the latent sequence as a spatiotemporal tensor $Z \in \mathbb{R}^{C \times T \times H' \times W'}$. Similar to stage 1, we apply the same noise to Z using two different sampled scales σ_s and σ_t, where $\sigma_t > \sigma_s \geq 0$. But the major difference from stage 1 is that we only apply full noise to the last frame so that the dynamics model F_ψ learns to predict the lower-noise last frame's latent given the action sequence and history context:

$$\widehat{Z}_{\sigma_s} = F_\psi(Z_{\sigma_t};\, \sigma_t, \sigma_s, a_{t-N:t-1})\,. \tag{4}$$

Figure 2 shows how we appended the noisy latent to history latents during inference time.

We train F_ψ with a weighted regression loss in latent space:

$$\mathcal{L}_{\text{dyn}}(\psi) = \mathbb{E}_{Z,\, \sigma_t > \sigma_s}\left[w(\sigma_t)\ \left\|\widehat{Z}_{\sigma_s} - Z_{\sigma_s}\right\|_2^2\right], \tag{5}$$

We instantiate F_ψ as a stack of 3D convolutional [78] blocks with FiLM modulation [79] and spatiotemporal attention to fully capture spatial temporal relations.

To enable long-horizon action-conditioned prediction, one important technique is injecting small noise to observation contexts so that the dynamics model F_ψ is robust to noisy contexts. During the online inference, models' prediction will be used as context for later steps, which is inevitably noisy. Therefore, model's robustness towards noisy context is very essential for stable long-horizon prediction.

3) Inference: During inference time, we predict future frames autoregressively. Given initial image o_0, we first obtain 2D latent $z_0 = E_\phi(o_0)$. As illustrated by Figure 2, we attach a noisy latent to history latents. Alongside action information, we denoise the last-frame noisy latent into a clean one (i.e., z_1). Then we attach the latest predicted latent to existing context to get the new history context (e.g., $z_{0:1}$). By repeating this process, we are able to autoregressively predict latents. Note that we discard the old latent once the context length is larger than a threshold so that the computation cost will not increase as horizon increases. Finally, the decoder D_θ takes in newly predicted latents z and renders new images $\hat{o}$.

C. Data Generation for Policy Training

Our Interactive World Simulator serves as a scalable surrogate for expert demonstration data collection, enabling the generation of high-quality synthetic demonstrations for imitation policy training without requiring access to physical robots.

Data collection is performed by initializing the simulator with an initial observation o_0, after which a human operator interacts with the simulator by issuing control commands through a teleoperation interface (e.g., keyboard input or low-cost kinematic devices). Given an action sequence $a_{0:T}$, the simulator autoregressively generates the corresponding observation sequence $o_{1:T}$, producing full demonstration trajectories $\{(o_t, a_t)\}_{t=0}^{T}$ that mirror real robot interaction data. The generated trajectories are directly compatible with standard imitation learning pipelines, including DP, ACT, and the π-series models, without requiring any modification to policy architectures or training procedures. This allows world simulator-generated data to be seamlessly mixed with real-world demonstrations or used as a standalone training source.

D. Policy Evaluation

Evaluating policies fairly is another longstanding challenge in robotics. Performing real-world evaluation is time-consuming and difficult to scale, as each trial requires manual resets and careful control of experimental conditions such as the initial object configuration. Our Interactive World Simulator enables scalable policy evaluation by allowing policies to interact directly with the simulated environment: the world model takes policy actions and predicts future frames, while the policy consumes the predicted frames to generate subsequent actions, mirroring real-world interaction. Compared to physical experiments, interacting within the world simulator enables controllable initial configurations, faster experiment resets, and safer large-scale evaluation. In the experiments section, we study whether policy performance in the world simulator faithfully predicts performance in the real world.

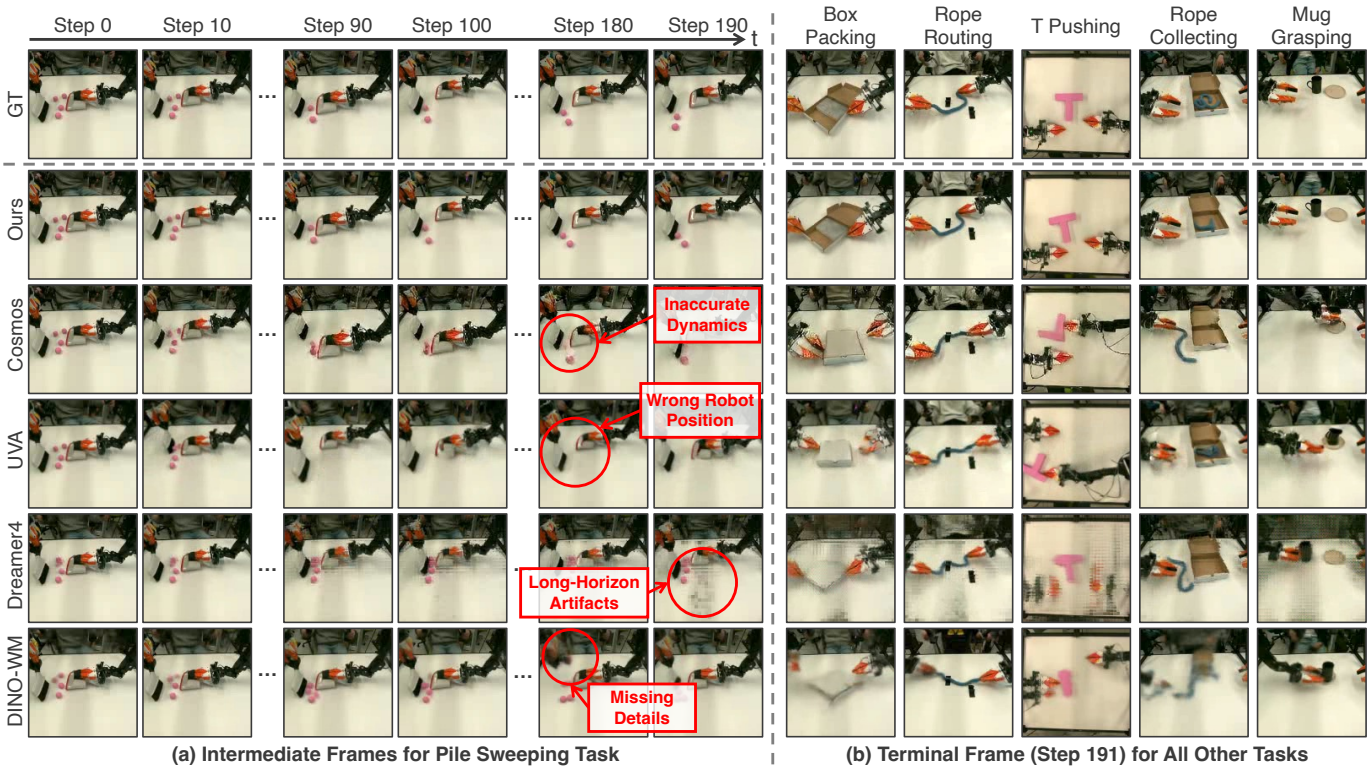

Fig. 3: **Qualitative Comparison of Long-Horizon Action-Conditioned Video Prediction.** We compare the proposed Interactive World Simulator with state-of-the-art action-conditioned video prediction models across multiple manipulation tasks. (a) Intermediate predicted frames for the `Pile Sweeping` task illustrate long-horizon rollout behavior, where baseline methods exhibit inaccurate dynamics, robot pose drift, accumulated artifacts, or loss of fine-grained details over time. (b) Terminal predicted frames for additional tasks, including `Box Packing`, `Rope Routing`, `T Pushing`, `Rope Collecting`, and `Mug Grasping`, highlight differences in visual fidelity and interaction consistency. In contrast to baseline methods, our model preserves coherent robot–object interactions and maintains stable predictions over extended horizons.

Appendix C
Experiment

In this section, we aim to study three questions: 1) How does our Interactive World Simulator perform compared to state-of-the-art world models in terms of realism, speed, and robustness? (Section C-B) 2) Can our Interactive World Simulator generate data with quality similar to real data for imitation learning? (Section C-C) 3) Can our Interactive World Simulator's policy evaluation faithfully represent real-world policy performance? (Section C-D)

A. World Model Instantiation

We set up one simulation task within MuJoCo [19] and six real-world tasks, including `Mug Grasping`, `Rope Routing`, `Rope Collecting`, `T Pushing`, `Box Packing`, and `Pile Sweeping`, which involve rigid objects, deformable objects, object piles, articulated objects, and the interactions among them. We performed all tasks using the ALOHA Bimanual Robot [12]. In simulation, we instantiate the T pushing task similarly to its real-world counterpart, as shown in Figure 1. We use a scripted policy to generate 10,000 episodes of random interactions to ensure sufficient data coverage. In the real world, we collect about 600 episodes of play data for each task, with each episode consisting of 200 steps. Real-world data collection for each task takes around six hours for a single person, which is a manageable scale for most research labs.

To make tasks interactive, we constrain part of the tasks' action space. For example, we limit motion of robot effector to bimanual plane motion for T pushing task. Additionally, our default image resolution is 128×128, which is sufficient for most of tasks.

Our model is lightweight. For example, the model for mug grasping task's size is 176.02 MB, making it easy to train and infer and keeps it very accessible for the broader research community. For stage 1 training, it usually takes around 6 hours on one H200 GPU, and around 12 hours for stage 2 training on one H200 GPU.

B. Video Baseline Comparisons

To evaluate the video prediction performance of the proposed Interactive World Simulator against prior action-conditioned world models, we conduct long-horizon action-conditioned video prediction comparisons across all six real tasks and one simulation task listed in Section C-A. We compare against representative state-of-the-art baselines, including Cosmos [6], UVA [8], Dreamer4 [16], and DINO-WM [3], under identical initial conditions and rollout horizons (192

Metric	Ours	DINO-WM [3]	UVA [8]	Dreamer4 [16]	Cosmos [6]
MSE ↓	**0.005** ± 0.005	0.028 ± 0.032	0.023 ± 0.015	0.012 ± 0.009	0.019 ± 0.010
LPIPS ↓	**0.051** ± 0.019	0.270 ± 0.093	0.272 ± 0.077	0.163 ± 0.052	0.224 ± 0.060
FID ↓	**63.50** ± 13.78	200.77 ± 79.02	142.55 ± 49.43	239.97 ± 45.75	200.74 ± 31.53
PSNR ↑	**25.82** ± 2.72	17.79 ± 2.84	17.87 ± 2.21	20.81 ± 2.21	18.91 ± 1.73
SSIM ↑	**0.831** ± 0.039	0.652 ± 0.059	0.650 ± 0.059	0.693 ± 0.045	0.647 ± 0.040
UIQI ↑	**0.960** ± 0.019	0.875 ± 0.029	0.884 ± 0.025	0.919 ± 0.024	0.883 ± 0.029
FVD ↓	**243.20** ± 103.58	1752.57 ± 805.56	2213.29 ± 525.48	1747.26 ± 248.08	799.34 ± 220.07

TABLE I: **Quantitative Comparison of Video Prediction Performance.** We quantitatively evaluate state-of-the-art action-conditioned video prediction models across a diverse set of manipulation tasks involving rigid objects, deformable objects, articulated objects, and object piles. The table reports results aggregated across all tasks; per-task results are provided in the supplementary material. Performance is measured using standard video prediction metrics, including MSE, LPIPS, FID, PSNR, SSIM, UIQI, and FVD, computed between predicted and ground-truth video rollouts. All results are reported for action-conditioned predictions over extended horizons (192 steps). The proposed Interactive World Simulator consistently outperforms prior world models across metrics, indicating improved visual fidelity, interaction consistency, and long-horizon stability, while maintaining efficient inference speed.

steps or 19.2 seconds). Prediction quality is assessed using standard reconstruction, perceptual, and temporal metrics such as MSE, PSNR, and FVD, computed between predicted and ground-truth video rollouts.

For baseline training, we follow the default settings in DINO-WM and Dreamer4 to train from scratch. We chose the community implementation version of Dreamer4 and followed their default training parameters. We followed the UVA training scripts to start the training from a pretrained VAE and another pretrained MAR model [80]. Cosmos provides guidelines on finetuning the Cosmos model. Therefore, we converted our data to the Cosmos data format and followed the default Cosmos finetuning settings.

Quantitative results in Table I show that the proposed method consistently outperforms prior methods across metrics, indicating improved visual fidelity and long-horizon temporal consistency. Table I aggregates results across all tasks, and the full result for each task is attached in the appendix.

Qualitative comparisons in Figure 3 further reveal different failure patterns of baseline methods such as robot pose drift, inaccurate object dynamics, missing details, or severe artifacts over long rollouts, whereas the proposed model maintains physically plausible interactions and stable predictions. In addition, the Interactive World Simulator achieves significantly higher inference efficiency, running at up to 15 FPS on a single GPU, enabling interactive long-horizon prediction that is impractical for many diffusion-based baselines. Our model could also infer stably for more than 10 minutes at 15 FPS, and more videos can be found on our project website.

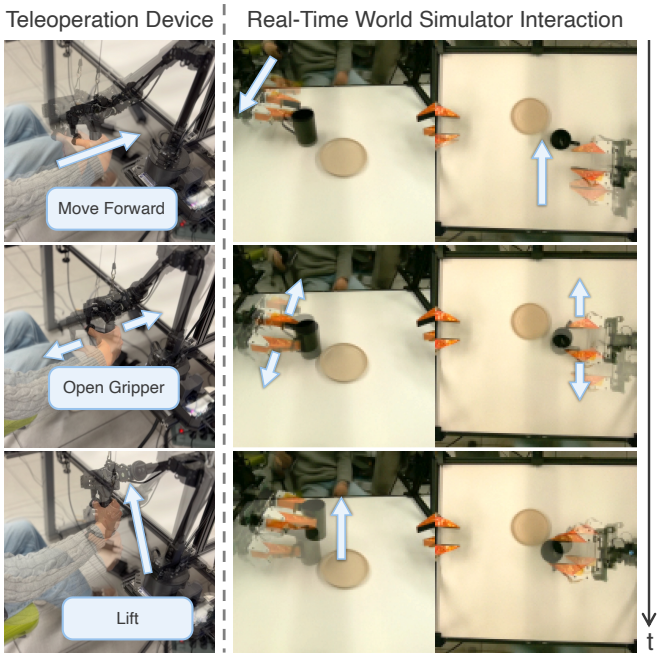

Fig. 4: **Teleoperation Interface for Real-Time Interaction with the World Simulator.** We build a kinematic teleoperation device that allows users to control robot actions through the Interactive World Simulator. The left column shows a user issuing commands through the teleoperation device (e.g., moving forward, opening the gripper, and lifting the end-effector). The right columns show the corresponding frames generated by the world simulator in real time. The simulator follows the teleoperation commands and produces consistent robot–object interactions.

C. Data Generation for Policy Training

To evaluate whether our world simulator can serve as a scalable data generator, we investigate whether synthetic demonstrations provide similar quality to real-world robot data for training imitation policies. We consider five manipulation tasks: T pushing in MuJoCo, T pushing in the real world, and three additional real-world tasks, pile sweeping, mug grasping, and rope routing. We benchmark four imitation learning policies: DP, ACT, π_0, and $\pi_{0.5}$. For each task, we use exactly 100 demonstration episodes to train imitation policies, regardless of the data source.

For fair and consistent policy evaluation, we define the task score for each task as follows across the whole paper, including Section C-D: (1) T pushing: maximum intersection-over-union between the T block's current pose and its target pose achieved within 600 steps; (2) Rope routing: number of clips through which the rope is inserted within 200 steps; (3) Mug grasping: one point for grasping the mug and one point for placing it on the plate within 200 steps; (4) Pile sweeping: number of pile pieces swept into the tray within 200

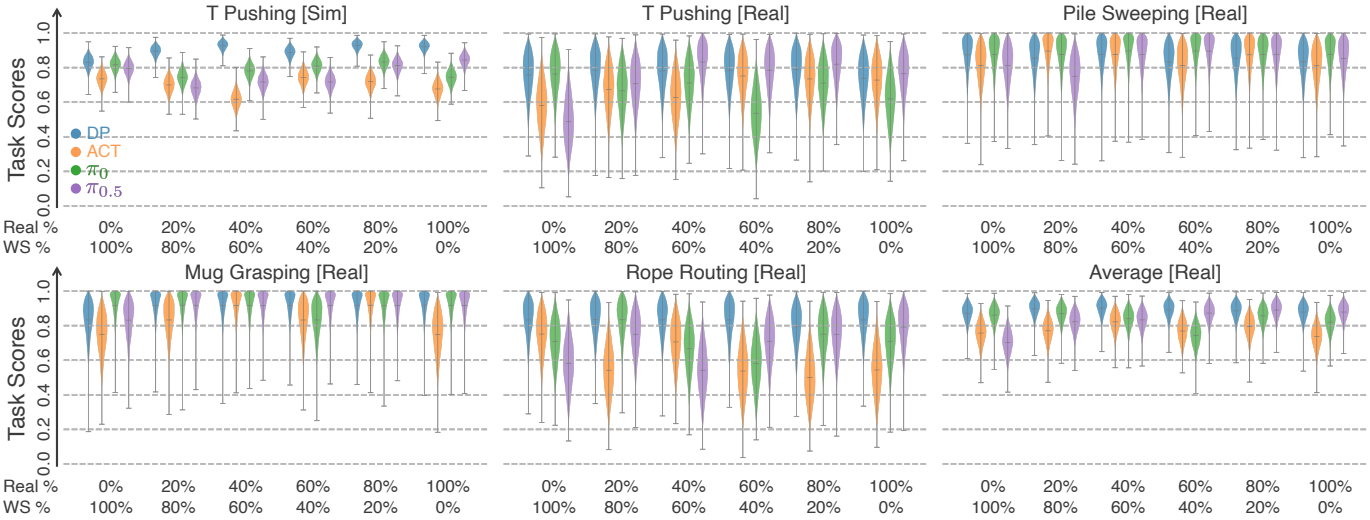

Fig. 5: **Imitation Policy Training with Mixtures of Real-World Data and Our World Simulator Data.** "WS %" represents the percentage of data from our world simulator, and "Real %" represents the percentage of data from the real world. From left to right, the mixtures range from 100% world simulator data to 100% real-world data. We use a total of 100 training episodes for each policy and data mixture. Performance among the policies trained on different data mixtures across different manipulation tasks is observed to be consistently high with comparable task scores, suggesting that data generated by our world simulator is comparable in quality to real-world demonstrations.

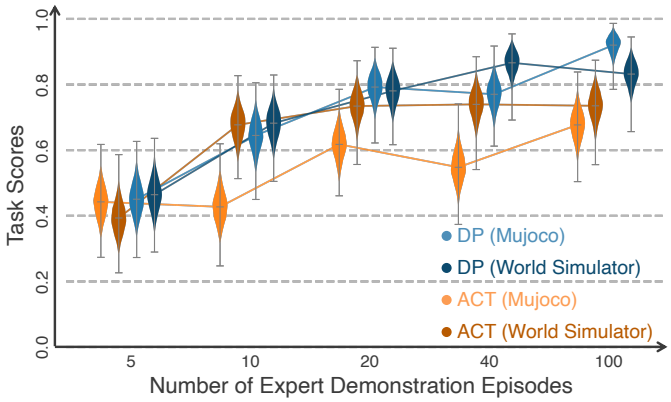

Fig. 6: **Effect of Data Scaling on Policy Performance**. We compare the performance of ACT and DP trained on datasets of varying sizes, ranging from 5 to 100 expert demonstration episodes. For each dataset size, policies are trained using either 100% MuJoCo data or 100% data generated from our world simulator. Policy performance improves consistently as the number of demonstrations increases across both data sources, suggesting that the scaling behavior of policy training is similar when using world simulator data.

steps. The rollout horizons for each task were chosen based on the average length of the expert demonstrations collected for the task plus some extra buffer of 100 steps to allow the policy rollout additional time to converge.

The user could use keyboard or kinematic devices to interact with the world simulator to collect expert demonstration data. Figure 4 showcases the teleoperation experiences using kinematic devices. On the left, it shows the process of a user controlling the teleoperation device and pretending to grasp the mug. The right process happens entirely in the world simulator and is very close to the real-world rendering. But in fact, there are no mugs or plates on the table. This example showcases the immersive experience of the world simulator interaction, allowing the user to collect high-quality expert demonstration data for training imitation learning policies.

To evaluate the quality of the simulator's data, we conduct two experiments. First, we construct training sets using varying mixtures of real-world data and world simulator-generated data with constant 100-episode dataset sizes. Then we train imitation learning policies with these data mixture ratios, ranging from 100% world simulator data to 100% real data. For each data point in MuJoCo, we evaluate the policy 100 times. For each data point in the real world, we evaluate the policy 10 times. Then, we obtain Figure 5, which represents task scores' Bayesian posteriors under a uniform Beta prior and the task scores we observe, following the practice of Large Behavior Models (LBM) analysis [15].

As shown in Figure 5, we observe consistent policy performance across the entire spectrum of data mixtures. Our analysis confirms that policies trained on 100% world simulator data achieve performance levels comparable to those trained on 100% real world expert data. For DP, the average task scores remain extremely stable, with a score of 87.9% using 100% simulator data compared to 90.3% using 100% real data. Similarly, ACT demonstrates strong parity, achieving 76.2% using simulator data versus 73.6% using real data. While $\pi_{0.5}$ shows a general trend of improvement with increasing real data, rising from 73.1% to 88.8%, it still achieves robust performance among the variety of tasks and initial configurations evaluated across simulation and real environments. Notably, policies trained on 100% world simulator data perform comparably to those trained on an equivalent volume of real-robot expert data. This suggests that our simulator can generate data

with quality similar to that of real-world demonstrations.

In the second set of experiments, we investigate whether the performance of imitation learning policies improves at a similar rate as additional data are introduced from different domains (real world vs. world simulator). We evaluate ACT and DP trained on datasets of varying sizes, ranging from 5 to 100 episodes. For each dataset size, we construct two training sets: one entirely from MuJoCo and one entirely from our world simulator. All policies are evaluated over 100 random trials. As shown in Figure 6, both policies exhibit consistent performance improvements as the training set size increases. Notably, data generated by our world simulator yield performance competitive with the MuJoCo baselines across all dataset sizes. In the low-data regime, policies trained on world simulator data successfully capture the task structure, matching the success rate of MuJoCo-trained policies. As the dataset scales to 100 episodes, the performance remains comparable. These results suggest that the scaling behavior of robotic policy training holds similarly within our world simulator, validating it as an effective tool for scalable data generation.

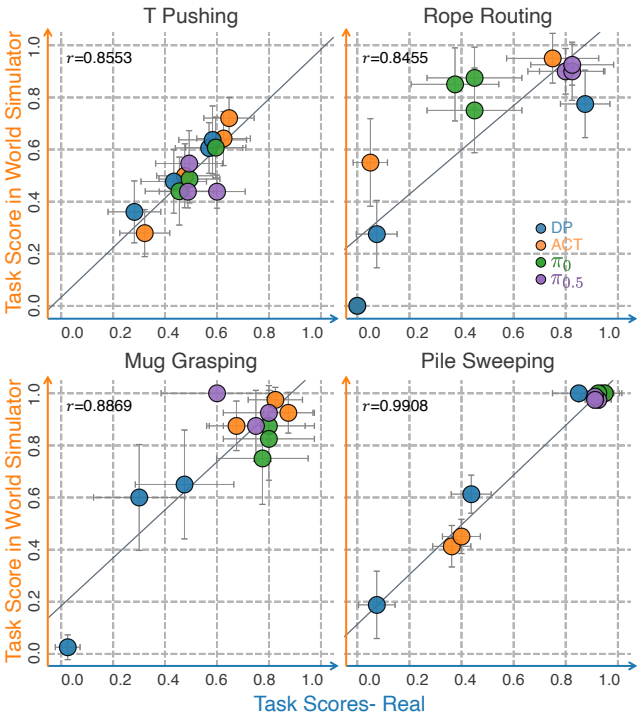

Fig. 7: **Correlation Between Policy Performance in Our World Simulator and the Real World.** Real-world task performance of imitation policies is plotted against their corresponding performance in our world simulator across multiple manipulation tasks. Each point corresponds to a trained policy evaluated under identical settings in both environments. Strong positive correlations are observed, indicating that evaluation within our world simulator closely reflects relative policy performance in the real world.

D. Sim-to-Real Correlation for Faithful Policy Evaluation

We investigate whether our world simulator can serve as a proxy for real-world policy evaluation and faithfully represent real-world performance. For this to hold, policy performance in the world simulator must be positively correlated with performance in the real world under identical initial conditions.

We train four imitation policies, DP, ACT, π_0, and $\pi_{0.5}$, on real-world data and evaluate the final and intermediate checkpoints of the policies on four manipulation tasks: T pushing, rope routing, mug grasping, and pile sweeping. For each task, we sample 20 initial configurations from the world simulator's training distribution and deploy each policy in both the world simulator and the real world.

Figure 7 shows the normalized scores averaged over the 20 configurations, with the x-axis representing real-world performance and the y-axis representing the world simulator performance. The error bars indicate Clopper-Pearson confidence intervals. Across all four tasks, we observe strong positive correlations between simulator and real-world performance. For tasks other than T pushing, the fitted lines exhibit a slight positive bias, indicating that policies tend to achieve slightly higher scores in the world simulator than in the real world. Despite the sim-to-real gap, this bias does not diminish the simulator's utility: if one policy substantially outperforms another in the world simulator, our results suggest that it is likely to do so in the real world as well. This property is useful in practice: one can rely on the world simulator to select high-quality candidate policies without the need for real-world experiments, thus reducing development cost and iteration cycles.