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/

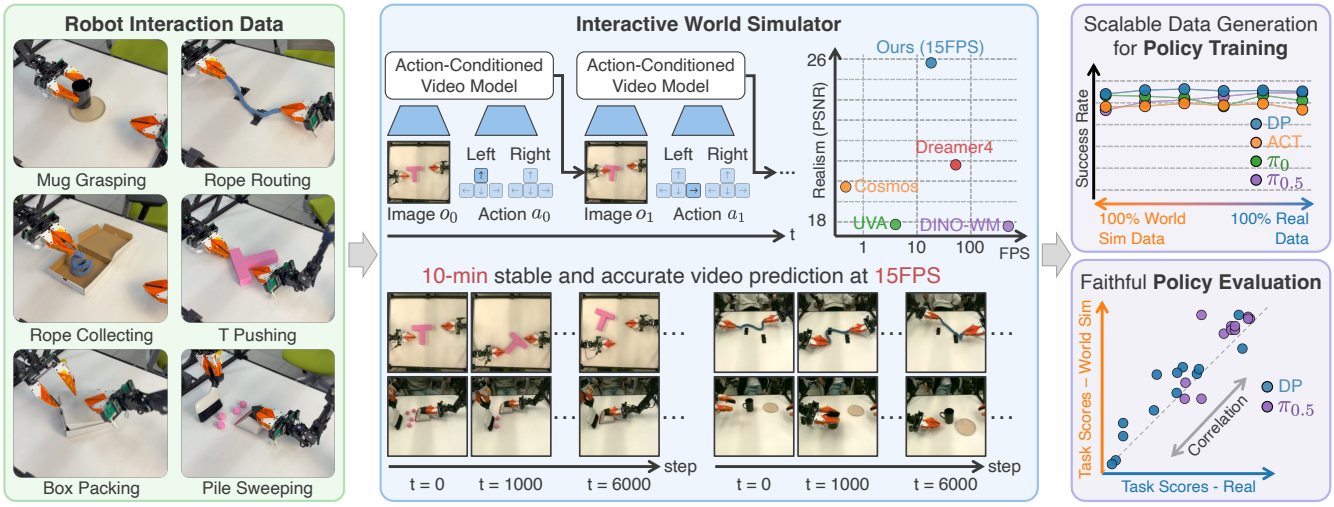

Fig. 1: **Overview of the Interactive World Simulator.** Given real-world robot interaction datasets (left), we train action-conditioned video models that capture complex physical dynamics and support stable long-horizon interactions (middle). The resulting world models achieve stable video prediction 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

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.

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.

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 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.

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. Evaluating policies fairly is another longstanding challenge in robotics. Performing real-world evaluation is time-consuming

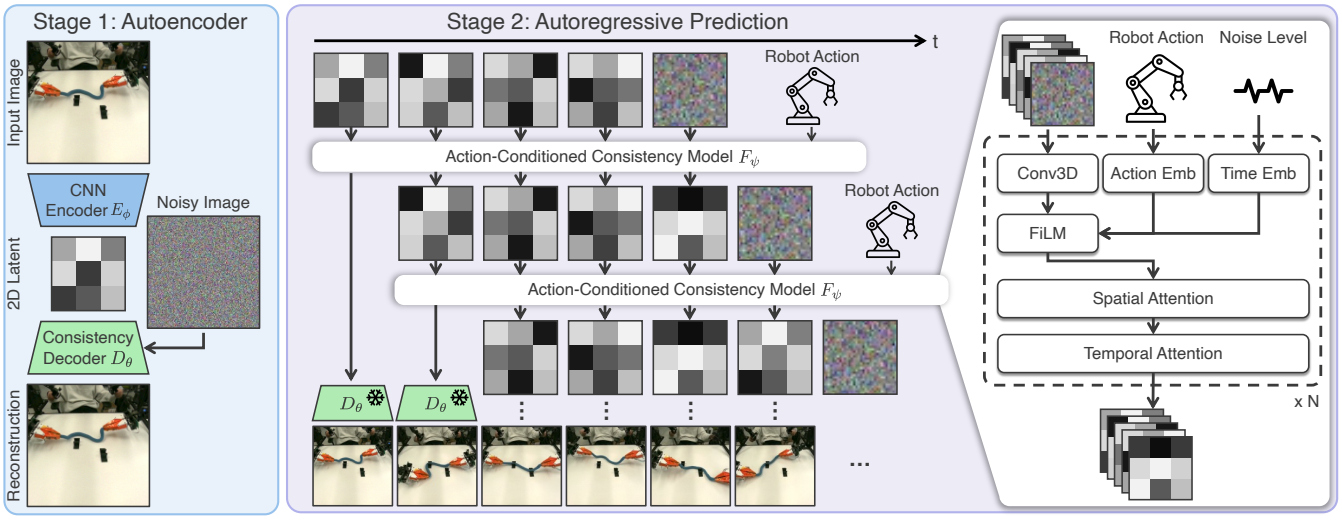

Fig. 2: **Method Overview.** Model training proceeds in two stages. In Stage 1, we train an autoencoder that maps RGB observations into compact 2D latent representations using a CNN encoder E_ϕ and a consistency-model decoder D_θ, enabling efficient and high-fidelity reconstruction. In Stage 2, we freeze 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.

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.

III. Experiment

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, deformable, articulated, piled, and multi-object interactions. 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, and use a scripted policy to generate 10,000 random-interaction episodes for data coverage. In the real world, we collect about 600 play-data episodes per 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 video prediction, we compare the proposed Interactive World Simulator against prior action-conditioned world models across all six real tasks and one simulation task listed in Section III-A. We compare against 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 and Policy Evaluation

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, pile sweeping, mug grasping, and rope routing. We benchmark DP, ACT, π_0, and $\pi_{0.5}$, training each imitation policy with exactly 100 demonstration episodes per task.

Our analysis confirms that policies trained on 100% world simulator data achieve performance comparable to those trained on 100% real-world expert data. DP remains stable, with an average score of 87.9% using simulator data compared to 90.3% using real data. Similarly, ACT achieves 76.2% using simulator data versus 73.6% using real data. This suggests that our simulator can generate data with quality similar to that of real-world demonstrations.

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. Across all four tasks, we observe strong positive correlations between simulator and real-world performance.

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.

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.