CVPR #30

Efficient Sim-to-Real Transfer of World-Action Models from Synthetic Priors

Anonymous CVPR submission

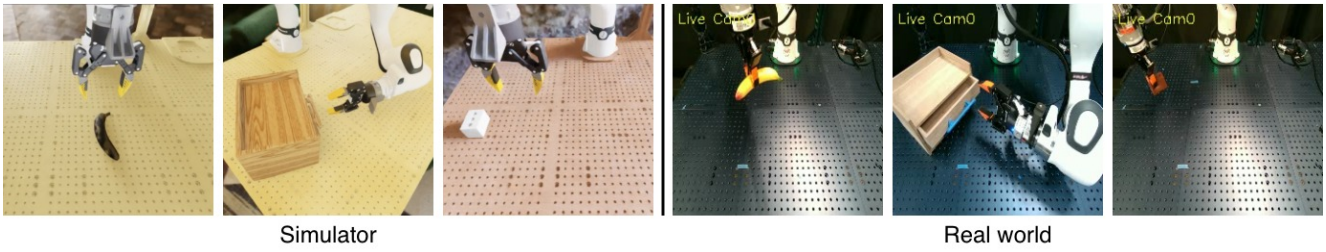

Figure 1. Zero-shot RGB sim-to-real transfer across tasks: *lift banana*, *lift brick* and *open drawer*.

Abstract

Bridging the sim-to-real gap is a core challenge in deploying learned manipulation policies. World-action models show remarkable ability in modeling environment dynamics. These models jointly predict future visual observations and robot actions within a unified generative framework. We argue this modeling capability is essential for effective sim-to-real transfer. To demonstrate this, we build upon Cosmos Policy, a video diffusion model adapted for visuomotor control. We construct simulation environments with extensive domain randomization, whereas demonstrations are then generated using the AnyTask motion planning pipeline. We evaluate our approach across object lifting, drawer opening, and pick-and-place tasks using only about 800 synthetic demonstrations per task. Ultimately, our policy achieves zero-shot RGB sim-to-real transfer on a Franka Research 3, attaining a 35% average success rate. To our knowledge, ***this represents the first successful sim-to-real transfer of a world-action model for robotic manipulation****.*

1. Motivation

Training manipulation policies in the real world is prohibitively expensive. While simulation offers a scalable alternative through parallelized data collection [8, 11], inherent visual and physical discrepancies, known as the sim-to-real gap, often lead to transfer failures [9, 10]. Current approaches struggle to bridge this gap efficiently. For instance, Vision-Language-Action (VLA) models like RT-2 [1] and OpenVLA [6] shows no sim-to-real effort, and architectures such as Diffusion Policy [2] decouple perception from action, limiting their ability to learn robust sim-to-real transferable behaviors.

We conject *world-action model*, a unified generative model that jointly predicts future visual observations and robot actions [3, 5, 7], is a promising direction. Rather than learning a narrow pixel-to-action mapping, world-action models must also forecast *what the world will look like next*, forcing them to internalize the environment dynamics. However, prior research on world-action models has been restricted to sim-to-sim or real-to-real deployment.

To investigate our assumption, in this paper we combine Cosmos Policy with extensive domain randomization and AnyTask's [4] automated demonstration generation, and present, to our knowledge, the *first successful sim-to-real transfer of a world-action model*.

2. Methodologies & Experiments

2.1. Cosmos Policy

Cosmos Policy [7] adapts a pretrained video diffusion model (Cosmos-Predict2) into a robot policy through single-stage post-training. In this framework, actions, future observations, and value estimates are seamlessly encoded as latent frames within a unified diffusion process.

2.2. Simulation and Demonstration Generation

We construct training environments using GPU-accelerated simulation [8] with comprehensive domain randomization: (1) *textures*: object surfaces, table materials, and backgrounds drawn from procedural and photographic libraries; (2) *camera poses*: perturbed translations and rotations of wrist and third-person cameras; (3) *lighting*: varied number,

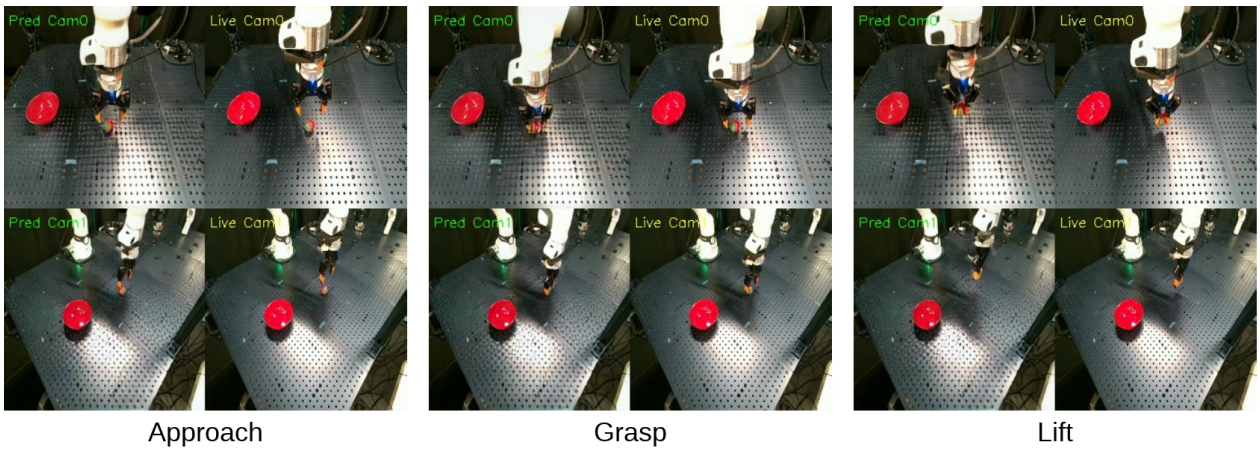

Figure 2. Predicted *vs*. live camera observations during *put strawberry into bowl* at each stage (approach, grasp, lift). The model learns accurate forward dynamics for real scenes despite training only in simulation.

position, intensity, and color temperature of scene lights; (4) *object placements*: uniformly sampled within the reachable workspace. This ensures the real world appears as just another variation within the training distribution [10].

We leverage AnyTask [4] to generate synthetic priors, i.e. expert demonstrations, via foundation-model-guided motion planning (ViPR) without human teleoperation across four tasks (*lift banana*, *lift brick*, *open drawer*, *put strawberry into bowl*) as Fig. 2, producing ∼**800 demos** per task (∼3,200 total) of RGB observations paired with end-effector action trajectories.

2.3. World-Action Model Training

The training of Cosmos Policy runs for **32** epochs on **40 H100** GPUs (∼**72** hours). As shown in Fig. 2, the model produces accurate future frame predictions matching real-world observations, confirming it has internalized forward dynamics despite training solely in simulation.

2.4. Zero-Shot Sim-to-Real Transfer Results

We deploy the trained policy directly on a real robot without any real-world fine-tuning or calibration beyond basic camera setup. Table 1 summarizes the success rates across the four tasks.

Table 1. Zero-shot sim-to-real success rate (%). Our world-action model uses only simulation demos; Diffusion Policy (DP) uses real demonstrations.

Method	**Banana**	**Brick**	**Drawer**	**Strawb.**	**Avg.**
DP w/ 10 real demos	0/10	0/10	2/10	0/10	**5%**
DP w/ 50 real demos	4/10	3/10	**3/10**	0/10	**25%**
Ours (800 sim)	**5/10**	**5/10**	2/10	**2/10**	**35%**

Our trained model achieves a **35%** average success rate using only ∼800 fully synthesized simulation demonstrations per task and *zero* real-world demonstrations, outperforming a Diffusion Policy [2] baseline trained on 50 real demonstrations per task (25%). This efficiency stems from pretrained video priors that reduce sample complexity, coupled with a joint action-video prediction objective that provides a self-supervised dynamics signal robust enough to transfer across visual domains. Notably, even without attempting to replicate the real-world testing environment's visual appearance during training, the model accurately predicts future observations.

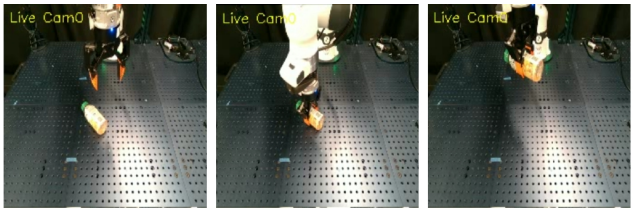

Figure 3. Out-of-distribution generalization: our policy lifts a novel bottle never seen during training.

Out-of-distribution generalization. We further evaluate lifting a novel bottle unseen during training (Fig. 3). The policy successfully approaches, grasps, and lifts it, showing the model acquires *generalizable visuomotor primitives* rather than memorizing task-specific trajectories.

3. Conclusion

We presented, to our knowledge, the first zero-shot sim-to-real transfer of a world-action model, achieving 35% success on four real-world tasks with only ∼800 fully automated simulation demos per task and zero real data. By jointly generating future video and actions, the model builds transferable physical understanding that generalizes across domains and to novel objects, establishing world-action models as a data-efficient paradigm for sim-to-real learning.

References

[1] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, 2023. 1

[2] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023. 1, 2

[3] Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[4] Ran Gong, Xiaohan Zhang, Jinghuan Shang, Maria Vittoria Minniti, Jigarkumar Patel, Valerio Pepe, Riedana Yan, Ahmet Gundogdu, Ivan Kapelyukh, Ali Abbas, Xiaoqiang Yan, Harsh Patel, Laura Herlant, and Karl Schmeckpeper. Anytask: an automated task and data generation framework for advancing sim-to-real policy learning. *arXiv preprint arXiv:2512.17853*, 2025. 1, 2

[5] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. 1

[6] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. 1

[7] Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, and Jinwei Gu. Cosmos policy: Fine-tuning video models for visuomotor control and planning. In *International Conference on Learning Representations*, 2026. 1

[8] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, and Gavriel State. Isaac Gym: High performance GPU-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021. 1

[9] OpenAI, Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, Jonas Schneider, Nikolas Tezak, Jerry Tworek, Peter Welinder, Lilian Weng, Qiming Yuan, Wojciech Zaremba, and Lei Zhang. Solving Rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019. 1

[10] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017. 1, 2

[11] Wenshuai Zhao, Jorge Peña Queralta, and Tomi Westerlund. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. *arXiv preprint arXiv:2009.13303*, 2020. 1