Encoder-Adapted Sim-to-Real Transfer of Simulation-Trained Diffusion Policies for Robot Manipulation

Chanhyuk Jung [1] **Dasom Ahn** [1] **Sungkeun Yoo** [2] **Byoung Chul Ko** [1]

Abstract

In this paper, we propose simDP, a sim-to-real transfer framework that enables diffusion policies trained in simulations to be deployed on real-world robots. The key idea is to reduce the sim-to-real gap by aligning the action and observation spaces between simulation and reality. This design allows the action decoder trained in simulation to be reused in the real-world with minimal modification. We evaluated simDP on object manipulation tasks derived from the MimicGen benchmark and show that a simulation-trained diffusion decoder, when combined with a real-world adapted observation encoder, achieves task-completion performance similar to and in some cases better than diffusion policies trained only on limited real-world data. These results suggest that reusing a simulation-trained action decoder with lightweight real-world encoder adaptation provides a practical and scalable solution for sim-to-real transfer in embodied robotic learning.

1. Introduction

Recent advances in large language models have shown that scaling model capacity together with high-quality data and improved training strategies can produce remarkable generalization ability across a wide range of tasks. A similar trend has emerged in multimodal learning, where vision-language models extend this paradigm by jointly processing textual and visual inputs and enabling reasoning over both modalities. Building on this progress, robotics has begun to adopt analogous formulations in the form of vision-language action (VLA) models, which aim to generate robot actions conditioned on language instructions and sensory observations.

Despite this promise, robotic learning remains fundamentally more difficult to scale than language or vision learning. Unlike text and image data, which can often be collected from large online corpora, robot manipulation data typically require teleoperation, physical hardware, carefully designed task environments, and repeated human intervention. As a result, the collection of large-scale real-world robotic datasets is expensive, slow, and difficult to parallelize (Open X-Embodiment Collaboration et al., 2023). However, in simulated environments, task configurations can be generated more easily, environment resets can be automated, and privileged state information can be accessed directly for supervision or evaluation. Therefore, simulation has become an increasingly important platform for training data-intensive visuomotor policies, including diffusion-based policies and emerging VLA robotic systems.

In this paper, we propose simDP, a sim-to-real transfer framework for diffusion policies (Chi et al., 2025). To bridge the sim-to-real gap, we unify the action-space and observation space between simulations and real-world environments. We evaluate simDP on object manipulation tasks derived from the MimicGen benchmark (Mandlekar et al., 2023) and test transfer performance on a physical robot. Experimental results show that our method achieves performance similar to, and in some cases better than, diffusion policies trained directly on limited real-world demonstrations. These findings indicate that aligning shared action representations and latent observation spaces is an effective strategy for sim-to-real transfer in diffusion-based robotic learning. Our contributions are summarized as follows:

- We propose simDP, a sim-to-real transfer framework for diffusion policies based on a shared action space and latent observation alignment.
- We show that end-effector-based actions enable the simulation-trained action decoder to be transferred across simulation and real-world robotic setups.
- We demonstrate that policies transfered with simDP achieves similar or superior performance compared

[1]Department of Computer Engineering, Keimyung University, Daegu, Republic of Korea [2]Department of Robot Engineering, Keimyung University, Daegu, Republic of Korea. Correspondence to: Sungkeun Yoo <skyoo@kmu.ac.kr>, Byoung Chul Ko <niceko@kmu.ac.kr>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

with diffusion policies trained on real-world data.

2. Method

Figure 1 shows an overview of simDP. First, we define a shared action representation that is valid in both simulation and real-world environments. Second, we decompose the policy into an observation encoder and an action decoder, and adapt only the real-world observation encoder while keeping the simulation-trained decoder fixed. This design allows the action-generation process to be learned from abundant simulation data and reused in the real-world through latent-space compatibility.

Instead of using robot-specific joint commands as actions, we represent each action around the end-effector. Let

$$a_t = [p_t, r_t, g_t], \tag{1}$$

where $p_t \in \mathbb{R}^3$ denotes the Cartesian position of the end effector, $r_t \in \mathbb{R}^6$ denotes the end-effector orientation represented using a continuous 6D rotation representation (Zhou et al., 2019), and $g_t \in \{0, 1\}$ denotes the binary gripper state. This representation is robot-agnostic in the sense that it describes task-relevant end-effector motion independently of the specific joint configuration of a given robot. The corresponding joint angles for each embodiment are recovered by inverse kinematics at execution time.

Let the observation at time t be denoted by o_t and consist of the visual input and proprioceptive state. Following DP (Chi et al., 2025), we predict an H-step future action chunk

$$a_{t:t+H-1} = (a_t, a_{t+1}, \ldots, a_{t+H-1}). \tag{2}$$

During training, a diffusion step $k \in \{1, \ldots, K\}$ and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ are sampled, and the clean action chunk is corrupted as

$$a^{(k)}_{t:t+H-1} = \sqrt{\bar{\alpha}_k}\, a_{t:t+H-1} + \sqrt{1 - \bar{\alpha}_k}\, \epsilon, \tag{3}$$

where $\bar{\alpha}_k$ is the cumulative noise schedule coefficient at step k. Given the simulation observation encoder f_{sim} and noise prediction network ϵ_ϕ, the simulation-stage training objective is

$$\mathcal{L}_{\text{sim}} = \mathbb{E}\left[\left\|\epsilon - \epsilon_\phi\Big(f_{\text{sim}}(o_t), a^{(k)}_{t:t+H-1}, k\Big)\right\|_2^2\right], \tag{4}$$

In our implementation, f_{sim} is a data-augmented ResNet-18 encoder (He et al., 2016) and ϵ_ϕ is a 1D U-Net (Ronneberger et al., 2015) following the standard DP architecture.

2.1. Real-World Observation Encoder Adaptation

Although the shared end-effector action representation reduces the embodiment gap, the visual observation distributions of the simulation and the real-world still differ significantly. Differences in texture, lighting, camera characteristics, and robot appearance prevent the simulation observation encoder from generalizing directly to real images. To address this issue, we replace the simulation encoder with a real-world observation encoder f_ψ and adapt only this encoder using real-world demonstrations, while keeping the simulation-trained noise prediction network fixed.

Importantly, our objective trains the real encoder to produce latent features that are compatible with the frozen decoder. Given real-world demonstrations, the real encoder is optimized using

$$\mathcal{L}_{\text{real}} = \mathbb{E}\left[\left\|\epsilon - \epsilon^*_\phi\Big(f_\psi(o^{\text{real}}_t), a^{(k),\text{real}}_{t:t+H-1}, k\Big)\right\|_2^2\right], \tag{5}$$

where ϵ^*_ϕ denotes the frozen simulation-trained decoder and only the encoder parameters ψ are updated. This objective encourages the real-world encoder to map observations into a decoder-compatible latent space, thereby enabling the simulation-trained action generator to operate on real-world inputs.

To make the shared action representation valid across domains, we also calibrate the simulation and real-world setups so that the workspace scale, camera viewpoint, and end-effector coordinate convention are relatively similar.

3. Experiments

This section presents the evaluation of simDP on three real-world manipulation tasks derived from MimicGen: *Stack Three*, *Three Piece Assembly*, and *Threading*. The main objective is to verify whether a simulation-trained DP can be effectively transferred to a physical robot through simDP. We further examine whether the proposed transfer strategy can outperform policies trained directly on limited real-world data. To provide a broader empirical comparison, we include several representative policy baselines: DP (Chi et al., 2025), flow matching policy (FMP) (Fang et al., 2026), and action chunking transformer (ACT) (Zhao et al., 2023).

3.1. Real-World Results

Table 1 reports the real-world normalized task completion (NTC) results on the three evaluation tasks. We reconstructed the MimicGen task environments on a real-world robot platform and evaluated all methods using the same task objectives as in the original MimicGen tasks. The main evaluation metric is NTC, measured as an average over 10 randomly initialized real-world rollouts. Because policies often complete only part of a task, we quantify progress using subtask completion milestones and normalize the resulting score to the range [0,1], where 1.0 indicates that all subtasks required for full task completion are successfully

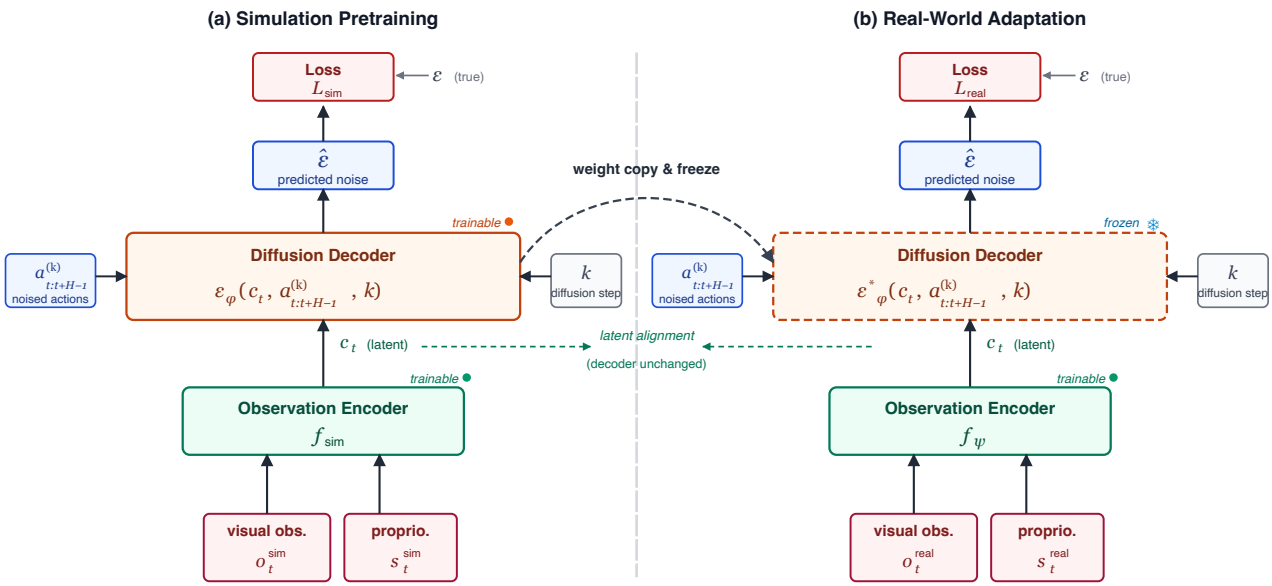

Figure 1. Overview of simDP. **Left:** In the simulation stage, the policy is trained using a shared end-effector action representation and a simulation observation encoder f_{sim}. **Right:** In the real-world adaptation stage, the diffusion decoder ϵ^*_ϕ is frozen and only the real observation encoder f_ψ is optimized so that its latent output becomes compatible with the frozen decoder. Here, $a^{(k)}_{t:t+H-1}$ denotes the noised action chunk at diffusion step k, and $\hat{\epsilon}$ denotes the predicted noise.

Table 1. Real-world normalized task completion scores (0–1) on the MimicGen-derived tasks. Each value is the average over 10 real-world rollouts. The best result in each column is shown in bold.

METHOD	STACK THREE	THREE PIECE ASSEMBLY	THREADING	AVERAGE
ACT (ZHAO ET AL., 2023)	0.75	0.35	0.70	0.60
DP (CHI ET AL., 2025)	0.75	0.30	0.80	0.62
FMP (FANG ET AL., 2026)	0.75	0.40	0.80	0.63
SIMDP (OURS)	**0.90**	**0.65**	**0.90**	**0.82**

achieved. All methods were evaluated with the same initial state and inference pipeline to ensure a fair comparison.

The results in Table 1 indicate that simDP consistently outperforms both the implicit policies based on generative models and the explicit policies based on the transformer architecture on all tasks. Compared with DP, the performance gain is relatively moderate on *Stack Three* and *Threading*, but increases on *Three Piece Assembly*. This trend suggests that the benefit of simulation pretraining becomes more pronounced as the task becomes more contact-rich. Overall, these results support the central claim of this paper: a simulation-trained diffusion decoder can be effectively reused in the real-world when the observation interface is properly adapted and the action space is shared across domains.

3.2. Sim-to-real Transfer Ablations

To verify the effectiveness of the proposed transfer mechanism, we compared simDP with four alternative strategies:

Table 2. Sim-to-real transfer comparison using normalized task completion scores (0–1). Each value is the average over 10 real-world rollouts.

TRAINING STRATEGY	STACK THREE	THREE PIECE ASSEMBLY	AVERAGE
ZERO-SHOT	0.00	0.00	0.00
FROM SCRATCH	0.75	0.30	0.53
SIMDP	**0.90**	**0.65**	**0.78**
SIMULATIONS	0.74	0.46	0.60

(1) zero-shot sim-to-real transfer, (2) training from scratch on real-world data, (3) the proposed decoder reuse with encoder adaptation, and (4) DP evaluated in simulation. The results are shown in Table 2. As expected, direct zero-shot transfer leads to catastrophic failure in which the model fails to understand the scene due to the huge sim-to-real gap in the observation space. Training DP from scratch enables the model to complete some subtasks, but because of the limited quantity and quality of the demonstrations, real-world performance is degraded. In contrast, simDP can

Table 3. Data-efficiency study with varying numbers of real-world demonstrations. Each entry is reported as the average over 10 real-world rollouts for *Stack Three* / *Three Piece Assembly*.

#SAMPLES	10	25	50
DP	0.20 / 0.10	0.45 / 0.25	0.75 / 0.30
SIMDP (OURS)	**0.55 / 0.35**	**0.85 / 0.60**	**0.90 / 0.65**

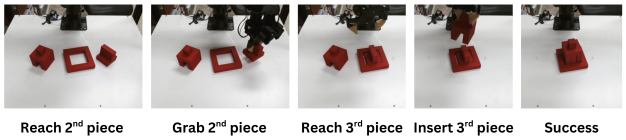

Figure 2. Representative real-world rollout of simDP on the *Three Piece Assembly* task. The sequence illustrates the successful execution of multi-stage manipulation, including grasping, alignment, insertion, and final assembly completion.

apply the skills learned from simulations into the real-world. This result suggests that utilizing the simulation-trained decoder is beneficial, since it already captures a strong action-generation prior learned from a large simulation dataset.

3.3. Real-World Data Efficiency

One of the main motivations of simDP is to reduce dependence on large-scale real-world demonstrations. To evaluate this aspect, we compared simDP and DP with varying numbers of real-world demonstrations. Table 3 shows that simDP outperforms DP trained from scratch when the same amount of demonstrations are used. The performance gap is especially large in thewlow-data regime, where the simulation-trained decoder provides a strong prior and the real-world encoder requires only modest supervision to adapt to the target domain. This behavior is desirable because the proposed method is intended for precisely the setting in which collecting real-world robot data is expensive and limited.

3.4. Qualitative Analysis

Figure 2 presents representative real-world rollouts of simDP on the *Three Piece Assembly* task. This qualitative sequence illustrates that the proposed policy can successfully execute multiple subtasks in the correct order, including grasping, alignment, insertion, and final assembly. This qualitative evidence complements the quantitative results by showing that the performance gain of simDP is associated with stable multi-stage behavior rather than isolated short-horizon successes.

4. Conclusion

In this paper, we proposed simDP, a sim-to-real transfer framework for diffusion-based robotic manipulation. The key idea of simDP is to reduce the sim-to-real gap from two perspectives: a shared action representation and real-world observation adaptation. Specifically, we define actions using end-effector pose and binary gripper state so that the same action space can be used in both simulation and real environments. In addition, we decompose the diffusion policy into an observation encoder and an action decoder, and adapt only the real-world encoder while keeping the simulation-trained decoder fixed. This design enables the policy to reuse the action-generation capability learned from scalable simulation data while requiring only limited real-world adaptation of the observation encoder. Evaluations on three MimicGen-derived real-world manipulation tasks suggest that effective sim-to-real transfer can be achieved without the need to retrain the full diffusion policy in the real-world.

5. Acknowledgement

This research was supported by Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government (MOTIE) (P0020536, HRD Program for Industrial Innovation) and the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education of the Republic of Korea (No. RS-2025-25399600).

References

Chi, C., Xu, Z., Feng, S., Cousineau, E., Du, Y., Burchfiel, B., Tedrake, R., and Song, S. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10–11): 1684–1704, 2025.

Fang, Y., Zhang, X., Cheng, H., Zang, X., Song, R., and Zhao, J. Flow policy: Generalizable visuomotor policy learning via flow matching. *IEEE/ASME Transactions on Mechatronics*, 31(1):140–150, 2026. doi: 10.1109/TMECH.2025.3586019.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Mandlekar, A., Nasiriany, S., Wen, B., Akinola, I., Narang, Y., Fan, L., Zhu, Y., and Fox, D. MimicGen: A data generation system for scalable robot learning using human demonstrations. *arXiv preprint arXiv:2310.17596*, 2023.

Open X-Embodiment Collaboration, O'Neill, A., Rehman, A., Maddukuri, A., et al. Open X-embodiment: Robotic learning datasets and RT-X models. *arXiv preprint arXiv:2310.08864*, 2023.

Ronneberger, O., Fischer, P., and Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241. Springer, 2015.

Zhao, T. Z., Kumar, V., Levine, S., and Finn, C. Learning fine-grained bimanual manipulation with low-cost hardware. In *Proceedings of Robotics: Science and Systems (RSS)*, Daegu, Republic of Korea, 2023.

Zhou, Y., Barnes, C., Lu, J., Yang, J., and Li, H. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5745–5753, 2019.