2026 26th International Conference on Control, Automation and Systems (ICCAS)
October 27-30, 2026. Sapporo, Japan

Enhancing Visuomotor Manipulation Control with a Predictive-Dynamics-Aware World-Model-Conditioned Diffusion Policy

Sangho An[1†], Geunha Kim[2†] and Jun Moon[1*]

[1]Department of Electrical Engineering, Hanyang University,
222 Wangsimni-ro, Seongdong-gu, Seoul 04763, Korea
[2]Department of Artificial Intelligence Systems, Hanyang University,
222 Wangsimni-ro, Seongdong-gu, Seoul 04763, Korea
(shan1005@hanyang.ac.kr, rlarmsgk2@hanyang.ac.kr, junmoon@hanyang.ac.kr)
[†] Equal contribution, * Corresponding author

Abstract: Recent visuomotor policies have achieved strong performance in robotic manipulation. Specifically, diffusion policy further improves action generation in various tasks. However, the diffusion policy mainly relies on image-based observation features and does not explicitly learn a dynamics-aware representation that summarizes temporal scene progression or object interactions. Therefore, we propose a world-model-conditioned diffusion policy (WM-DP) for visuomotor manipulation. WM-DP extends the original diffusion policy by concatenating a latent feature generated by a world model. Through this design, we study whether a simple early-stage world model conditioning structure can improve real-world manipulation performance without adding architectural complexity. We evaluate WM-DP on real-world pick-and-place tasks using objects with different physical properties. We also evaluate the method using a deformable object whose graspable regions and appearance change depending on its orientation. The results show that our model improves both success rate and trajectory quality under observed object conditions. Overall, our work suggests that even a GRU-based world model can provide useful cues for physical interaction and improve diffusion policy in visuomotor manipulation.

Keywords: Visuomotor manipulation, Diffusion model, World model

1. INTRODUCTION

Recent visuomotor control methods have shown impressive performance in robotic manipulation by directly mapping visual observations to continuous robot actions [1–3]. In particular, imitation-learning-based policies and diffusion-based policies have enabled accurate and smooth manipulation behaviors in a wide range of object-centric tasks [4, 8, 13, 14]. These results demonstrate the strong potential of visual policies for end-to-end robotic control, but they also reveal that reliable action prediction still requires an understanding of temporal scene evolution and robot–object interaction dynamics.

This remains challenging because the same visual observation can lead to different actions depending on object pose, contact timing, and the previous motion of the robot. Therefore, a visuomotor policy needs to understand the current scene while also using temporal context from recent history. This issue is also related to the multi-modal demonstrations, where similar scenes may correspond to different reaching directions, grasp poses, or recovery motions. While diffusion policy [8] alleviates the averaging problem of standard behavior cloning by generating action sequences through conditional denoising, its standard CNN-based formulation [5] still relies mainly on image features and does not explicitly encode temporal scene progression or recent action tendencies.

A single observation is insufficient to fully represent temporal information such as object motion, contact timing, and recent interaction history. To provide this missing information, we condition the diffusion policy on a latent feature produced by a world model. The world model receives recent observations and produces a latent representation that contains dynamics-related information. We do not use the world model as a learned simulator. It does not generate imagined rollouts or replace the diffusion policy structure. Instead, we use its latent feature only as an additional condition for the denoising network. This feature is concatenated with the base observation condition, allowing our model to generate actions using both visual information and recent dynamics cues.

We evaluate the proposed method on simple yet dynamics-sensitive manipulation tasks. From an object-centric perspective, we further demonstrate that the proposed method can generate effective grasping angles according to the object orientation. Across both tasks, the proposed method shows improved performance over the standard diffusion policy.

2. RELATED WORK

2.1 Diffusion Models

Diffusion models have been widely used for modeling complex and high-dimensional distributions. They were first popularized in image generation through denoising-based generation methods such as DDPM [6] and score-based models [7]. In diffusion models, a neural network is trained to remove noise step by step and recover a clean sample from a noisy input. Although diffusion models were originally developed for image generation [11], their denoising-based formulation has been broadly extended to other domains, ranging from diffusion transformer [9], diffusion-based language mod-

Copyright ©2026 ICROS

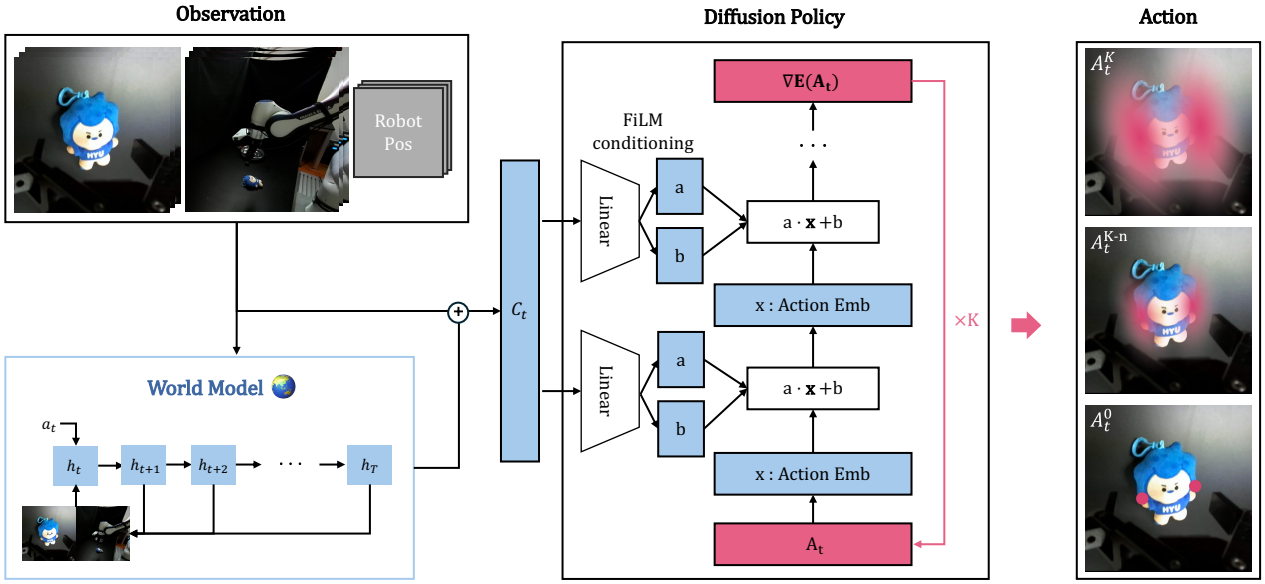


Fig. 1. **Structure of the world-model-conditioned diffusion policy.** The proposed WM-DP augments the original observation condition with the world model latent h_t^{wm}. The concatenated condition $\tilde{C}_t$ is then used by the denoising network to generate the action sequence A_t.

els [10], and AlphaFold [12] to diffusion-policy-based visuomotor control. When the sample is defined as a robot action trajectory, the same denoising process can be used for action generation.

Diffusion policy [8] applies this idea to visuomotor manipulation. Instead of directly predicting a single action, it generates a short sequence of future actions conditioned on recent observations. Only the first few actions are executed, and the policy replans after receiving new observations. This closed-loop action generation is effective for manipulation tasks where multiple valid action trajectories may exist from the same visual observation.

Our method follows the same diffusion-based action generation framework, but modifies the conditioning input. While the original diffusion policy mainly conditions the denoising network on observations, we additionally provide a hidden feature generated by a world model branch. This structure allows the policy to use both the current observation condition and an additional temporal feature for action generation.

2.2 World Models

World models aim to learn representations of dynamics from observations and actions [15]. Related ideas have also appeared in earlier model-based control methods, such as learning latent dynamics directly from raw images [16] and using probabilistic dynamics models for sample-efficient reinforcement learning [17]. Recently, this idea has been extended to broader areas such as embodied AI, autonomous systems, and video-based prediction, where temporal scene progression and physical interaction are important [22]. In robot manipulation, such representations can provide useful cues about object motion, contact, and recent interaction history.

A common way to model these temporal dynamics is to use a state-space model (SSM), where the hidden state summarizes the recent history of observations and actions. Recurrent state-space models (RSSMs) [18–21] further combine recurrent hidden states with latent variables, allowing the model to represent dynamics in a compact latent space [18]. Recently, this latent-dynamics formulation has been widely used in world-model-based reinforcement learning.

Some methods use world models as learned simulators for imagined rollouts or additional policy training. DiWA [24], for example, uses a world model to generate imagined trajectories for offline policy improvement. In contrast, our method does not use the world model as a simulator and does not perform reinforcement learning or policy fine-tuning. Instead, we use a world model as an auxiliary conditioning branch for diffusion policy, providing dynamics-aware cues to the denoising network.

3. METHODOLOGY

3.1 World Model Conditioning

In this work, we adopt an earlier version of the world model concept and integrate it into our diffusion policy framework. Instead of using a large-scale world foundation model or a complex dynamics model, we implement the world model branch with a simple GRU-based structure [23]. This design choice was made for two reasons. First, we wanted to examine whether the basic idea of world-model-based conditioning is effective for improving diffusion policy. Second, the model should remain lightweight so that it does not introduce a large computational burden during real-world robot execution.

The GRU-based world model receives the observation history and updates its hidden state over time. The result-

ing hidden feature is used as a compact latent representation of recent dynamics. Rather than using the world model to generate imagined rollouts or reconstruct future observations, we directly use this latent feature as an additional conditioning input for the denoising network. Therefore, the proposed method can provide dynamics-aware cues to the policy while maintaining a simple and practical architecture.

3.2 Network Architecture

The overall pipeline of the proposed WM-DP is shown in Fig. 1. At each time step, the robot receives visual observations from the camera inputs and robot proprioceptive states, including the gripper state. These inputs are used as the main condition for diffusion policy. Each image observation is resized to a fixed resolution of $224 \times 224 \times 3$ before being passed to the policy.

In parallel, the recent observation history is given to the GRU-based world model branch. The world model updates its hidden state and produces a latent feature h_t^{wm}, which represents temporal and dynamics-related information from recent observations. This feature is then concatenated with the original observation feature to form the augmented condition $\tilde{C}_t$.

The augmented condition is defined as:

$$h_t^{\mathrm{wm}} = f_{\mathrm{wm}}(O_t),$$

$$\tilde{C}_t = [O_t^{\mathrm{obs}}; h_t^{\mathrm{wm}}], \quad p_\theta(A_t \mid \tilde{C}_t).$$

The denoising network receives $\tilde{C}_t$ and generates an action sequence A_t. As in the original diffusion policy, only the first few actions are executed on the robot, and the policy replans after receiving the next observation. Through this pipeline, WM-DP preserves the visual conditioning structure of diffusion policy while additionally using the temporal dynamics cue from the world model.

Following the original diffusion policy, we also use FiLM conditioning [25]. FiLM modulates intermediate feature maps by applying condition-dependent scale and shift parameters. Specifically, given a feature representation x, FiLM generates two condition-dependent parameters, γ and β, and transforms the feature as:

$$\mathrm{FiLM(x)} = \gamma \odot \mathrm{x} + \beta.$$

In our network, the augmented condition $\tilde{C}_t$, which includes both the current observation feature and the world model feature, is used to provide conditioning information to the denoising network through FiLM layers. This allows the denoising network to adapt its action generation process according to the current observation, temporal dynamics cue, and diffusion step.

4. EXPERIMENTS

We conducted pick-and-place experiments using two objects with different material characteristics. This setup was designed to examine manipulation behavior under both rigid and deformable object conditions. The experimental setup is summarized in Table 4, and we also evaluated the model under unseen object-orientation cases.

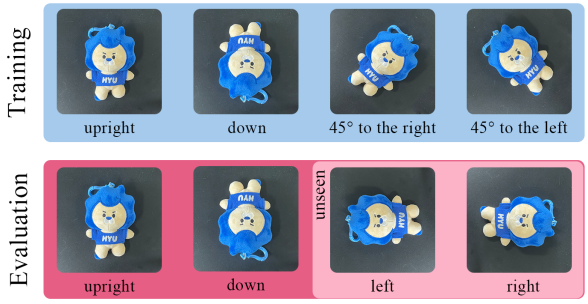


Fig. 2. HYlion orientation setup for object-centric grasping evaluation. The upper row presents the orientations included in the training demonstrations, and the lower row presents the orientations used for evaluation, including unseen left- and right-rotated cases.

4.1 Task Setup

We selected pick-and-place as the experimental task because it involves physical dynamics and 3D spatial interactions, as shown in Fig. 3. Through this task, we examine whether world model conditioning can help the policy capture real-world dynamics.

There were two target objects used for pick-and-place: a red block and a HYlion doll. Each object has different physical properties. If the gripper is not well aligned with a red block, it tends to slide off to the side as the gripper closes. This means that the red block highly requires precise gripper alignment.

In contrast, the doll has different characteristics from the red block. Since the doll is a deformable object, the robot does not always need to be precisely aligned to achieve a successful grasp. In some cases, it can successfully grasp the target by holding the loop on its head, the end of its foot, or by pinching part of its body. This makes it more tolerant to grasping errors compared to the red block. However, the doll also has an orientation-dependent appearance because its visible shape changes depending on its rotation. Therefore, it was used to evaluate whether the policy can generalize to diverse object orientations in real-world tasks.

In addition, we adopted a contact-and-release setting, where the robot had to grasp the object, move it over the hurdle, place it inside the basket, and release the gripper only after the object made contact with the bottom surface. Therefore, trials were considered failures if the object was grasped too shallowly or if the lifting height was insufficient to clear the hurdle. This setting made the task more challenging by requiring both stable grasping and precise object placement.

4.2 Results

We evaluated both policies under identical conditions. The main evaluation metrics were inference time, success rate, and performance under unseen 90° rotated cases.

Table 1. Inference time comparison on the manipulation task.

Method	Inference time (ms)
Diffusion Policy	183.6 ± 14.8
WM-DP (Ours)	253.6 ± 19.2

Table 2. Success rate on the red block pick-and-place task. The error term is calculated as the standard error.

Method	Success rate(%)
Diffusion Policy	45.0 ± 7.9
WM-DP (Ours)	72.5 ± 7.1

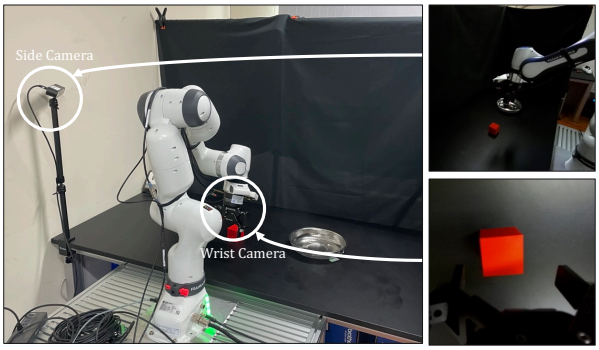
Inference time. We measured the inference time of each method. Inference time is largely determined by the iterative denoising process used in diffusion models. For a fair comparison, we used the same number of denoising steps, 80, for both diffusion policy and our model during evaluation. For each method, we conducted 10 trials across 4 different seeds. The results are summarized in Table 1. These results indicate that WM-DP introduces only a slight increase in execution time while providing a substantial improvement in task performance.

Success rate. Table 2 shows that our model achieved a success rate of 72.5%, corresponding to a 27.5 percentage-point improvement over the diffusion policy baseline. This substantial improvement can be attributed to the predictive dynamics provided by world model conditioning, which helped the policy handle the challenging task conditions intentionally designed in our experiment. To further examine this result, we analyze the observed failure patterns and compare the actual trajectories generated by each method.

For the red block, we observed a characteristic failure pattern. Although the baseline was often able to approach the object, the grasping point was not sufficiently deep. As a result, the object was blocked by the hurdle of the target basket, leading to task failure. Another major failure case was inaccurate approach motion. When the gripper was not properly aligned with the red block, the object tended to either slip away during gripper closing or be pushed aside as the gripper approached it.

In contrast, our model reduced these failure patterns and consistently achieved deeper, more stable grasps. It also showed better robustness when the object was placed in more challenging regions of the workspace. In several trials, our model even performed a re-grasping motion, adjusting the grasp before picking up the object. While the baseline sometimes dropped the object during transport due to early gripper opening, our model maintained a stable grasp until the object contacted the basket. These results suggest that our model provides more robust grasping and transport behavior in our tasks.

Object-centric grasping control. For HYlion, which has different physical characteristics from the red block,



(a) Real-World Setup (b) Image Observations

Fig. 3. **Real-world setup.** The experimental setup was configured as summarized in Table 4. Two camera views were used: a side-view camera and a wrist camera.

the success rate in the upright orientation was slightly higher than that of the red block pick-and-place task. This result is notable because the target demonstrations were not concentrated only on the upright orientation; instead, the training dataset included mixed orientations, such as upright, inverted, and approximately 45-degree rotated poses. In other words, even though the number of upright-only demonstrations for HYlion was relatively limited, the policy still achieved a higher success rate than in the red block task.

The deformability of the target object also affected the behavior of the baseline. In several successful trials, the baseline grasped the target through non-standard grasping points. These cases indicate that the doll can tolerate a wider range of grasping points than the red block. Table 3 summarizes the success rates for HYlion under different object orientations. Our model achieved an average success rate of 54.2%, while the diffusion policy baseline achieved 41.7%.

Our model outperformed the baseline in the upright, inverted, and left-rotated cases, improving the success rate by 30.0, 22.5, and 5.0 percentage points, respectively. This suggests that world model conditioning provides useful predictive cues for orientations included or partially represented in the demonstrations. However, performance decreased in the right-rotated case from 45.0% to 10.0%, likely because the demonstrations did not include end-effector rotations larger than 45 degrees, limiting generalization to complete 90-degree alignment.

Overall, these results indicate that our model improves performance when the physical interactions are within or near the range observed during training. At the same time, the degraded performance under the unseen 90-degree rotated condition suggests that the world model feature was actively involved in action generation, but its effectiveness was limited outside the range of physical dynamics observed during training.

4.3 Training

Dataset. We collected a total of 72 pick-and-place demonstrations, consisting of 36 demonstrations for the

Table 3. Success rate comparison for object-orientation-aware grasping control with HYlion. All values are reported in percentage.

Method	Up	Down	Left	Right	Avg.
Diffusion Policy	47.5	37.5	35.0	45.0	41.7
WM-DP (Ours)	77.5	60.0	40.0	10.0	54.2

red block and 36 demonstrations for HYlion, as shown in Fig. 2. For HYlion, the training data included demonstrations under multiple orientations: the canonical upright orientation, an orientation rotated by approximately 45 degrees from the upright pose, and an inverted orientation. To evaluate generalization performance of the policies to unseen object orientations, we additionally tested the trained policy on cases where the object was rotated 90 degrees to the left or right, which were not included in the demonstration dataset.

Training condition. Training for both diffusion policy and world-model-conditioned diffusion policy was conducted under the exactly same training conditions. Both methods were trained for 75 epochs with a batch size of 16 and an observation horizon of 2. By keeping the training conditions identical, we ensured a fair comparison between the two methods and evaluated their respective performance under the same experimental setting.

The overall experimental results can be found in `https://youtu.be/dVcs-mAfopQ`.

5. CONCLUSION

We designed a world-model-conditioned diffusion policy (WM-DP), where the policy is conditioned not only on the base observation but also on the latent feature generated by the world model.

The experimental results show that WM-DP achieved a higher success rate than the DP-only baseline under the same training and evaluation conditions. This improvement was not only observed in the quantitative success rate, but also in the qualitative behavior of the robot.

Our model produced more stable grasping and transport behavior than diffusion policy, suggesting that world model conditioning helps generate more reliable action sequences. However, the lower performance in the unseen 90-degree rotated HYlion condition also indicates an important limitation. Because this condition was not sufficiently represented during training, the world model feature affected action generation but did not provide reliable generalization outside the demonstrated range of physical interactions.

Another important finding is that all training and evaluation settings were controlled to be identical for DP-only and WM-DP. Therefore, the performance improvement was achieved without increasing the amount of training data or changing the training conditions. This indicates that world model conditioning can be integrated into diffusion policy as a lightweight structural extension while still providing meaningful performance gains.

Table 4. Real-world experimental setup.

Component	Specification
Manipulator	Franka Emika Panda
Gripper	ROBOTIQ Adaptive Gripper
Side-view camera	ZED 2
Wrist-mounted camera	ZED Mini
Task	Pick-and-place manipulation
Objects	Red block, HYlion (Hanyang Univ. mascot)

ACKNOWLEDGEMENT

This work was supported in part by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Climate, Energy & Environment (MCEE) of the Republic of Korea (No. RS-2024-00422103, EV Smart Charging Platform Innovation Research Center), in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korean government (Ministry of Science and ICT, MSIT) (RS-2020-II201373), in part by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (RS-2024-00397400), in part by the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. RS-2024-00421129), and in part by the Technology Innovation Program (20018112) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea).

REFERENCES

[1] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *arXiv preprint arXiv:1504.00702*, 2015.

[2] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, and S. Levine, "QT-Opt: Scalable deep reinforcement learning for vision-based robotic manipulation," *arXiv preprint arXiv:1806.10293*, 2018.

[3] A. Brohan et al., "RT-1: Robotics Transformer for real-world control at scale," *Robotics: Science and Systems*, 2023.

[4] Y. Go, Y. Oh, S. Kim, and J. Moon, "GASP: Generative Adversarial State Policy and Multi-modal Sensing Framework for Dexterous Manipulation," *International Journal of Control, Automation, and Systems*, 2026.

[5] K. He, X. Zhang, S. Ren, and J. Sun,"Deep residual learning for image recognition,"in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[6] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, 2020.

[7] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*, 2021.

[8] C. Chi et al., "Diffusion Policy: Visuomotor Policy Learning via Action Diffusion," *arXiv preprint arXiv:2303.04137*, 2024.

[9] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[10] S. Nie, F. Zhu, Z. You, X. Zhang, J. Ou, J. Hu, J. Zhou, Y. Lin, J.-R. Wen, and C. Li, "Large language diffusion models," *arXiv preprint arXiv:2502.09992*, 2025.

[11] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[12] J. Abramson et al., "Accurate structure prediction of biomolecular interactions with AlphaFold 3," *Nature*, vol. 630, no. 8016, pp. 493–500, 2024.

[13] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low cost hardware," in *Robotics: Science and Systems*, 2023.

[14] Y. Zhu, P. Stone, and Y. Zhu, "VIOLA: Imitation learning for vision-based manipulation with object proposal priors," in *Conference on Robot Learning*, 2023.

[15] D. Ha and J. Schmidhuber, "World Models," *arXiv preprint arXiv:1803.10122*, 2018.

[16] M. Watter, J. Springenberg, J. Boedecker, and M. Riedmiller, "Embed to control: A locally linear latent dynamics model for control from raw images," in *Advances in Neural Information Processing Systems*, 2015.

[17] K. Chua, R. Calandra, R. McAllister, and S. Levine, "Deep reinforcement learning in a handful of trials using probabilistic dynamics models," in *Advances in Neural Information Processing Systems*, 2018.

[18] D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson, "Learning latent dynamics for planning from pixels," in *International Conference on Machine Learning*, 2019.

[19] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, "Dream to Control: Learning Behaviors by Latent Imagination," in *International Conference on Learning Representations*, 2020.

[20] D. Hafner, T. Lillicrap, M. Norouzi, and J. Ba, "Mastering Atari with discrete world models," *arXiv preprint arXiv:2010.02193*, 2020.

[21] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap, "Mastering diverse domains through world models," *arXiv preprint arXiv:2301.04104*, 2023.

[22] C. Bodnar et al., "A foundation model for the Earth system," *Nature*, vol. 641, no. 8065, pp. 1180–1187, 2025.

[23] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[24] A. L. Chandra, I. Nematollahi, C. Huang, T. Welschehold, W. Burgard, and A. Valada, "DiWA: Diffusion Policy Adaptation with World Models," *arXiv preprint arXiv:2508.03645*, 2025.

[25] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.