DISTRIBUTION RECOVERY IN COMPACT DIFFUSION WORLD MODELS VIA CONDITIONED FRAME INTERPOLATION

Sam Gijsen & Kerstin Ritter
Department of Psychiatry and Psychotherapy, Charité - Universitätsmedizin Berlin, Berlin, Germany
Hertie Institute for AI in Brain Health, University of Tübingen, Tübingen, Germany
sam.gijsen@charite.de

ABSTRACT

This early proof-of-concept explores addressing distribution drift in diffusion-based world models without requiring massive model scale or constrained environments. We explore a dual-purpose training approach where models learn both autoregressive world generation and frame interpolation capabilities. This is combined with an out-of-distribution detection mechanism that, upon detecting drift or degradation, samples appropriate target frames and conditions the model to interpolate toward them, effectively pulling generation back into the learned distribution. We demonstrate this approach's potential through initial experiments and discuss practical considerations for target frame sampling and interpolation training. This early work presents an alternative path toward enabling longer world exploration with smaller models.

1 INTRODUCTION

World models based on diffusion processes have shown remarkable potential for generating coherent environmental sequences (Yang et al., 2024; He et al., 2025). However, these models often suffer from distribution drift during extended generation, where the quality and consistency of generated frames gradually deteriorate. Recent work has demonstrated impressive results in addressing this challenge through different approaches: Genie leverages model scaling (Bruce et al., 2024), while other work focuses on constrained environments such as a single video game level (Alonso et al., 2024), or combines both strategies as demonstrated by Valevski et al. (2025). However, these solutions prevent low-cost world exploration.

We explore an alternative approach that enables smaller diffusion world models to allow longer exploration through a dual-purpose training strategy. Our method combines traditional autoregressive world generation with frame interpolation capabilities (Lyu et al., 2024), allowing the model to detect and recover from distribution drift by conditioning on appropriately sampled target frames. This approach effectively creates "checkpoints" in the learned distribution that the model can navigate toward when generation quality degrades.

Using an early proof-of-concept implementation in a 3D game environment, we demonstrate that this method can significantly extend the coherent generation horizon of a compact diffusion world model. Our approach requires neither the scale of contemporary large models nor the strict constraints of specialized environments, suggesting a promising direction for efficient and robust world modeling.

2 METHOD

Our approach builds on the diffusion-based world modeling framework introduced by Alonso et al. (2024), which leverages score-based diffusion models to generate high-fidelity environment observations for reinforcement learning. Specifically, we adopt their DIAMOND (DIffusion As a Model Of eNvironment Dreams) methodology, employing the Elucidating Diffusion Models (EDM) formulation from Karras et al. (2022) to enhance training stability and sampling efficiency. This involves

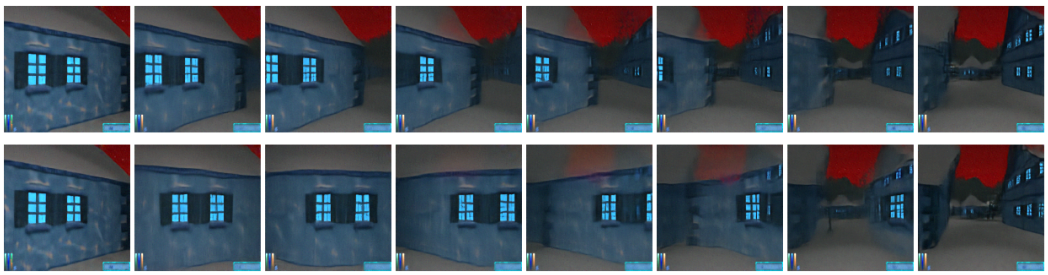

Figure 1: Subsets of generated frames by the model during interpolation towards a target frame (left-to-right) under different inputs: turn right (top), turn left (bottom).

a UNet-based denoiser conditioned on past observations and actions, trained to reverse a noising process as described by Song et al. (2021). The EDM framework uses network preconditioning to normalize input and output variances, enabling high-quality generation with minimal denoising steps. We extend this framework with a dual-purpose training strategy that combines autoregressive prediction and frame interpolation, as detailed below.

We train the UNet-based denoiser from scratch (330M parameters) that conditions on 4 previous frames and action inputs (forward, backward, left, right, turn right, turn left, jump) at 10 FPS. We extend this framework with a dual-purpose training strategy.

During training, we use a batch with probability $p_i = 0.25$ to train the model to interpolate towards a future frame over m=10 steps. We create $m + 1$ embedding buckets to indicate the interpolation step (or no interpolation for autoregressive generation). With probability 1-p_i, we perform standard autoregressive training with action conditioning over 10 steps, with the future frame target dropped out.

To prevent the model from defaulting to simply predicting the target frame at each step, we: (1) Weight sequences with significant frame differences more heavily in the loss. This addresses cases where the random agent gets stuck (e.g., walking into a wall), which would otherwise reinforce the local minimum of predicting identical frames. (2) Provide intermediate targets by interpolating between the current frame at time t_i and t_m, with increasing weight on t_m as interpolation progresses. As the frame at t_i is generated, we make interpolation more dynamic by incorporating action inputs and allowing some control during this process. Here we use simple image-space interpolation, as initial experiments with interpolation between VAE latent representations showed no qualitative differences, possibly due to the simplistic environments. (3) Train for quadratic rather than linear interpolation to promote adherence to autoregressive sampling and action conditioning during initial frames, creating more interesting dynamics and preventing immediate convergence to the final target frame.

During inference, we maintain a buffer of frames sampled uniformly from the dataset. When drift is detected through frame similarity metrics, we select a similar frame from this buffer as an interpolation target, effectively pulling generation back into the learned distribution. For these examples, we simply rely on the frame-by-frame MSE.

We train for two days on an NVIDIA GeForce 3090 after collecting 9 hours of random agent gameplay footage from the videogame The Elder Scrolls II: Daggerfall, using a 30x30 resolution denoiser with a separate 51M parameter upsampling model for visualization at 150x150.

3 RESULTS

We demonstrate our approach using an undertrained model to investigate the effectiveness of interpolation-based recovery. Training for two days on a single 3090 GPU yields a model capable of basic action-conditioned autoregressive generation, though prone to distribution drift.

Figure 1 illustrates how action conditioning influences the interpolation process. Given the same starting point and target frame but different action inputs (turn left vs. turn right), the model produces

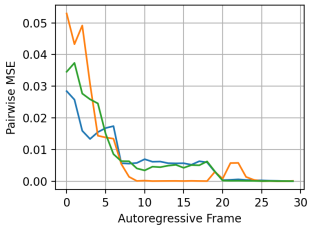

Figure 2: Pairwise MSE between consecutive frames for unsuccessful example autoregressive sequences. Each colored line corresponds to one sequence.

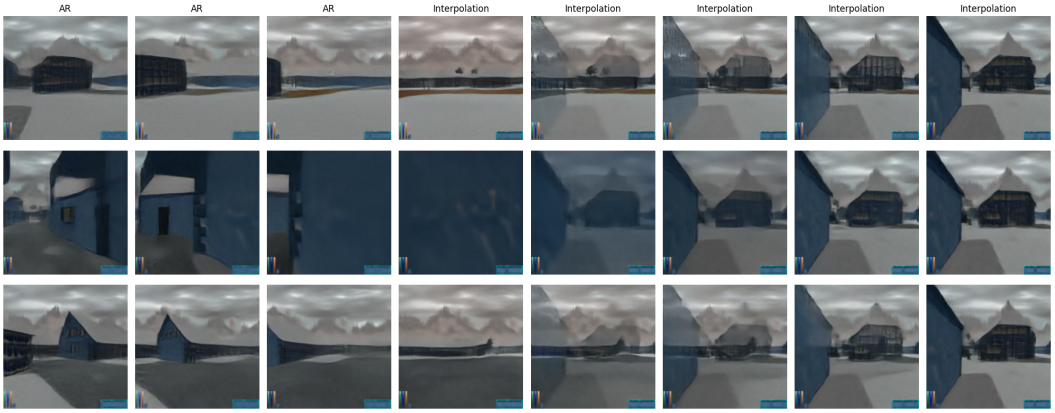

Figure 3: Autoregressive generation (left-to-right; frames are not consecutive) starts in-distribution and is seen to drift. Once generation breaks down and no significant changes are observed between frames, interpolation is started by providing a target frame. Following interpolation towards the target frame (rightmost image), stable autoregressive sampling can resume.

distinct intermediate frames while still reaching the target, suggesting it can maintain meaningful action conditioning during interpolation.

Figure 2 shows the mean squared error (MSE) between consecutive frames during three unsuccessful runs of autoregressive generation, where low MSE indicates model drift, at which point consecutive frames do not significantly differ anymore. Figure 3 demonstrates our recovery mechanism: when generation quality degrades and the model gets 'stuck' without meaningful changes between frames, interpolation toward a selected target frame successfully returns the model to the learned distribution. At this point, the frames used for conditioning are in distribution, which enables continued autoregressive generation.

Our proof-of-concept implementation has several limitations. The current frame selection strategy is simplistic, using basic MSE similarity. The interpolation process could benefit from more sophisticated transitions, potentially guided by the trajectory of player actions. Additionally, earlier detection or even predicting of impending distribution drift could enable more natural recovery sequences.

4 Discussion

This early work explores a simple approach to extending the generation horizon of compact diffusion world models. By combining autoregressive generation with frame interpolation capabilities, we demonstrate the potential for models to recover from distribution drift without requiring massive scale or environmental constraints.

Several directions for future work emerge from this proof-of-concept. The frame selection strategy could be expanded to consider player trajectories, selecting target frames that maintain contextual

consistency. More sophisticated interpolation mechanisms could enable smoother transitions. Finally, leveraging the denoiser's latent representations might enable earlier prediction of distribution drift, allowing more proactive and natural course corrections.

While our approach is preliminary, it suggests that explicit distribution recovery mechanisms might offer an alternative path toward stable world models that are more accessible to resource-constrained applications.

REFERENCES

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.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

Haoran He, Yang Zhang, Liang Lin, Zhongwen Xu, and Ling Pan. Pre-trained video generative models as world simulators. *arXiv preprint arXiv:2502.07825*, 2025.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Zonglin Lyu, Ming Li, Jianbo Jiao, and Chen Chen. Frame interpolation with consecutive brownian bridge diffusion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3449–3458, 2024.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=P8pqeEkn1H.

Sherry Yang, Yilun Du, Seyed Kamyar Seyed Ghasemipour, Jonathan Tompson, Leslie Pack Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=sFyTZEqmUY.