Playable World Models: Real-Time Diffusion-Based Video Generation for Physics-Consistent Snooker

Moin Arz Mattar

Abstract

Recent advances in world modeling suggest that generative video models may move beyond passive prediction and begin to function as interactive simulators. Rather than rendering frames via explicit physics engines, these models learn environment dynamics directly from data, conditioned on observations and actions. This raises a fundamental question: *can learned generative models preserve the minimal physical structure required for real-time interaction?*

In this project, we explore this question by constructing a fully playable snooker game driven entirely by real-time video generation. Using diffusion-based world models conditioned on player actions, we investigate whether such models can implicitly learn and maintain core physical behaviors—such as collision consistency and trajectory stability—well enough to act as a lightweight game engine. Our work evaluates the boundary between visually plausible prediction and physically usable interaction, offering insight into the feasibility of learned simulators for games, robotics, and embodied AI.

1 Introduction and Motivation

Traditional game engines rely on hand-crafted physics solvers and explicit symbolic state updates. While these systems are reliable, they require extensive manual design and are tightly coupled to specific environments. In contrast, recent work most notably from Google DeepMind demonstrates that large generative models can learn environment dynamics directly from sequences of observations and actions, enabling a new class of *data-driven world models*.

If diffusion-based models can reliably preserve simple physics in an interactive setting, this would represent a significant step toward learned simulators capable of generalizing beyond manually specified rules.

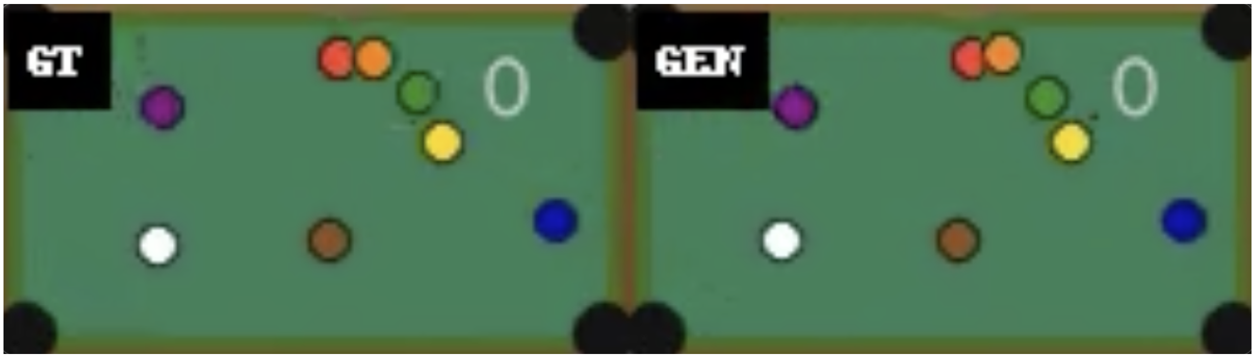

Figure 1: Enter Caption

Such models could support imagination-driven planning, embodied reasoning, and scalable environment simulation.

We focus on snooker as a controlled but unforgiving testbed. Although visually simple, snooker exhibits strict physical constraints:

- Small errors in collision handling or velocity propagation compound rapidly.
- Object interactions are frequent and highly structured.
- User actions (cue direction and force) have precise, interpretable effects.

These properties make snooker an ideal environment for evaluating whether generative video models can transition from *plausible rollouts* to *usable interactive dynamics*.

2 Technical Background

Our approach builds on three recent advances in representation learning and generative modeling.

Vision Transformers (ViT) provide a flexible framework for learning spatial representations from image patches, enabling object-centric reasoning over balls, table geometry, and relative spatial relationships.

Diffusion models for video generation iteratively refine predictions over time, offering improved temporal coherence and stability compared to autoregressive methods—properties that are critical for long-horizon interaction.

Diffusion Forcing extends diffusion models to the interactive setting by conditioning future frames on control actions, allowing the system to respond dynamically to user input rather than merely extrapolating observed motion.

Together, these components form the basis for a fully learned, action-conditioned world model operating directly in pixel space.

3 Proposed Method

We train a diffusion-based world model that takes as input:

- A short history of rendered video frames,
- A player action specifying cue direction and force,
- And outputs future video frames representing the resulting game evolution.

Importantly, the model does not explicitly encode latent physical state variables such as positions, velocities, or collision events. Instead, all dynamics are learned implicitly through video prediction conditioned on actions.

At inference time, the model operates in a closed loop, generating frames in real time and accepting new actions from the player, effectively functioning as a learned game engine.

4 Evaluation and Goals

We evaluate whether the learned world model preserves essential physical structure by analyzing:

- **Trajectory consistency**: smoothness and predictability of ball motion across time.
- **Collision behavior**: qualitative preservation of momentum transfer and relative motion.
- **Interaction stability**: robustness under repeated user interventions.

Our primary goal is not photorealism, but *behavioral consistency*. We seek to identify whether diffusion-based video models can support sustained, physically meaningful interaction rather than short-term visual plausibility.

5 Related Work

Our work draws inspiration from Vision Transformers [1], diffusion-based sequence modeling [2], recent large-scale world models from Google DeepMind [3], diffusion approaches to world modeling [4], and diffusion forcing for action-conditioned generation [5].

References

[1] Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv:2010.11929*

[2] Ho et al. Diffusion Models for Sequence Modeling. *arXiv:2212.09748*

[3] Google DeepMind. Scalable World Models. *arXiv:2408.14837*

[4] Diffusion for World Modeling. *arXiv:2405.12399*

[5] Diffusion Forcing. *arXiv:2407.01392*