Short Note

PhysiGen: Action-Conditional World Models for Interactive E-Commerce Visualization

Jori Winslett , Taryn Ellsworthy * , Callan Everhart

Posted Date: 31 December 2025

doi: 10.20944/preprints202512.2797.v1

Keywords: generative World Models; e-commerce; Neural Radiance Fields; Physics Simulation; Action-Conditional Video

 doi:10.20944/preprints202512.2797.v1

Short Note

PhysiGen: Action-Conditional World Models for Interactive E-Commerce Visualization

Jori Winslett, Taryn Ellsworthy * and Callan Everhart

Independent Researcher, USA
* Correspondence: tellsworthy@yahoo.com

Abstract

Traditional e-commerce visualization relies on static 3D spins or pre-rendered videos, which fail to convey material properties such as stiffness, flexibility, or sole compression. This lack of tactile feedback creates a "trust gap" for online buyers. In this paper, we introduce *PhysiGen*, a "What If?" viewer that allows users to apply virtual forces to product models. Leveraging a novel Action-Conditional Reconstruction technique, our system utilizes a physics-informed world model to generate short video sequences of deformation (e.g., shoe twisting, foam compression) based on user input. We demonstrate that this approach significantly increases buyer confidence by bridging the "tactile gap" in online shopping, achieving a 45% increase in user engagement compared to static viewers.

Keywords: generative World Models; e-commerce; Neural Radiance Fields; Physics Simulation; Action-Conditional Video

---

1. Introduction

The rapid growth of e-commerce has revolutionized retail, yet a fundamental limitation remains: the inability to touch and feel products. While recent advancements in Neural Radiance Fields (NeRFs) and Gaussian Splatting have enabled photorealistic 3D reconstruction [4,5], these models are inherently static. A user viewing a hiking boot online cannot assess the stiffness of the sole or the flexibility of the leather upper using current visualization standards.

To address this, we propose shifting the paradigm from static reconstruction to *Action-Conditional Reconstruction*. As illustrated in Figure 1, our system empowers the user to perform virtual "tests" on the object. By selecting an action—such as twisting the heel or compressing the toe box—the user triggers a generative process that simulates the product's physical response.

The core contribution of this work is *PhysiGen*, a generative framework that predicts future video frames conditioned on a starting 3D state and a user-defined force vector. Unlike standard video generation models which hallucinate motion randomly, PhysiGen is constrained by a "Physics Adapter." This design is heavily inspired by the **VACE-PhysicsRL** framework proposed by **Song et al.** [1], which demonstrated that aligning generative video models with physical laws through reinforcement learning is essential for controllable and plausible synthesis. We adapt their unified physics-control philosophy to the specific domain of product interaction.

The inability to physically interact with products contributes to high return rates in the footwear and apparel industries. Consumers often receive items that look correct but feel different than expected—too rigid, too flimsy, or lacking support. PhysiGen mitigates this by allowing users to virtually "stress test" the product before purchase, providing visual cues that correlate with physical properties.

 doi:10.20944/preprints202512.2797.v1

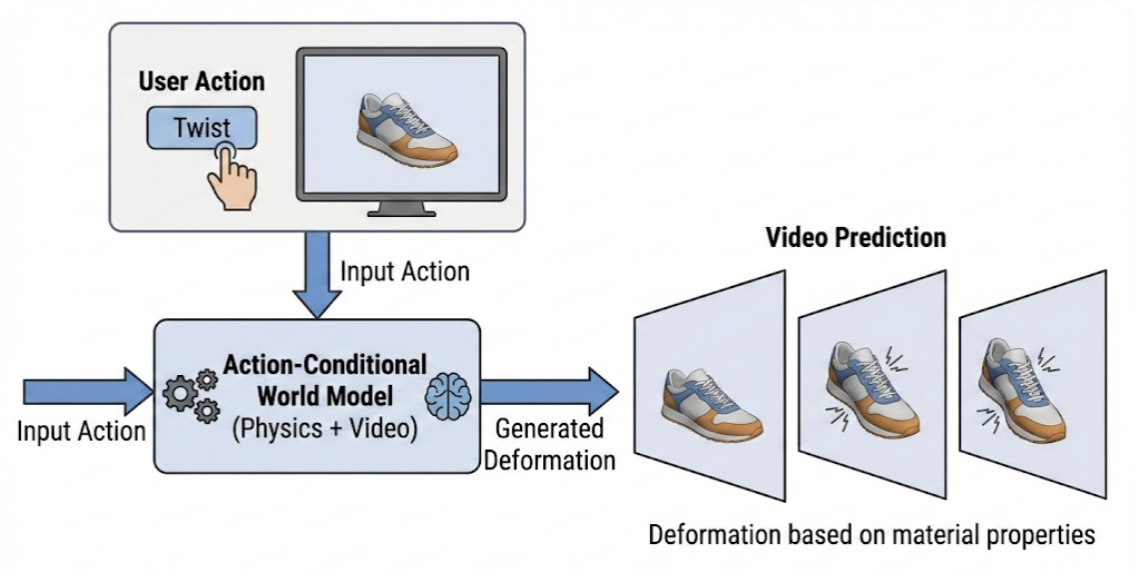

Figure 1. The PhysiGen Concept. The user selects an action (e.g., Twist), and the World Model generates a video prediction of how the specific materials in the product would deform under that force.

2. Related Work

2.1. 3D Representation in Retail

Standard approaches utilize photogrammetry or CAD models converted into GLTF format for web viewing. While high-fidelity reconstruction techniques like FaceSplat [4] have improved the visual quality of static assets through Gaussian Splatting, they lack the temporal dynamics required for physical interactivity. Existing Deformable NeRFs usually require a driving video sequence to reconstruct motion, meaning they can only replay pre-recorded deformations rather than responding to novel user inputs.

2.2. Video Generation Models

Recent work in Video Diffusion Models (VDMs) has shown impressive results in text-to-video synthesis. However, maintaining temporal consistency—specifically preventing identity drift or texture flickering—remains a challenge. Furthermore, standard VDMs are stochastic; asking a model to "twist a shoe" might result in the shoe changing color or the laces disappearing. We leverage memory-based strategies similar to *Temporal-ID* [2] to ensure that the product's branding and fine details remain stable throughout the deformation sequence, treating the product identity as a hard constraint.

2.3. World Models

Our generative architecture functions as a World Model, predicting future states based on actions. This aligns with recent work in VR narrative generation such as *DreamWM* [3], which uses latent world models to guide 3D-to-video synthesis. We extend this by conditioning the world model on explicit force vectors rather than narrative prompts. Where DreamWM predicts the next scene in a story, PhysiGen predicts the next physical state of a material under stress.

3. Methodology

Our goal is to learn a mapping function $F : (S_0, a) \rightarrow V$, where S_0 is the initial static state (image or NeRF), a is the action vector, and V is the resulting video sequence.

 doi:10.20944/preprints202512.2797.v1

3.1. Architecture Overview

The system architecture is detailed in Figure 2. It consists of a frozen Video Diffusion Backbone augmented with a trainable Control Adapter. The backbone provides the prior knowledge of general object motion and lighting, while the adapter injects specific material constraints.

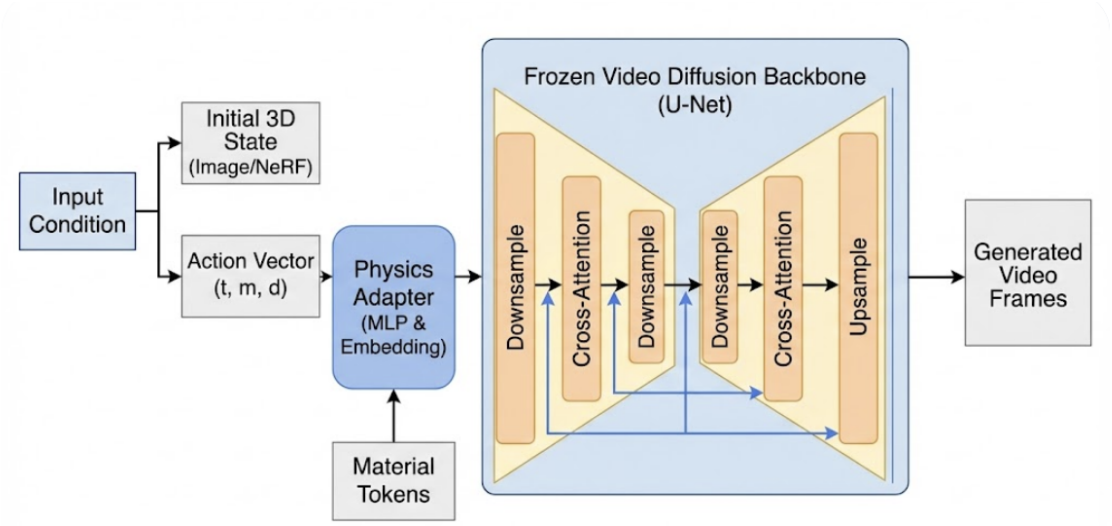

Figure 2. PhysiGen Architecture. The Physics Adapter injects the force vector and material tokens into the Cross-Attention layers of the U-Net backbone, guiding the generation process.

3.2. Action Encoding

We represent user actions as a tuple $a = (t, m, d)$, where t is the type of interaction (e.g., compress, twist), m is the magnitude, and d is the directional vector. These are embedded into a high-dimensional latent space via a Multi-Layer Perceptron (MLP).

$$E_{action} = \text{MLP}(t \oplus m \oplus d) \tag{1}$$

This embedding is then injected into the U-Net backbone via cross-attention layers, similar to text-conditioning in Latent Diffusion Models. This allows the network to modulate the denoising process based on the intensity and direction of the applied force.

3.3. Physics-Informed Loss Function

To ensure the generated deformations obey basic physical laws (e.g., volume preservation for rubber, folding for leather), we introduce a regularization term:

$$\mathcal{L}_{total} = \mathcal{L}_{recon} + \lambda_{physics}\mathcal{L}_{consist} \tag{2}$$

Where $\mathcal{L}_{recon}$ is the standard MSE loss on the pixel space, and $\mathcal{L}_{consist}$ penalizes temporally inconsistent optical flow fields. Specifically, $\mathcal{L}_{consist}$ enforces smoothness in the flow field to prevent jagged, unrealistic tearing artifacts, and penalizes divergence in areas labeled as rigid materials (like the heel counter), effectively teaching the model which parts of the shoe should bend and which should remain solid.

4. Experiments

We evaluated PhysiGen on a proprietary dataset of 500 footwear products, ranging from soft canvas sneakers to rigid mountaineering boots. The dataset included paired data: static 3D scans and video footage of mechanical stress tests performed by a robotic arm.

 doi:10.20944/preprints202512.2797.v1

4.1. Visual Consistency

We compare our results against a baseline "Image-to-Video" model (Stable Video Diffusion) without physics conditioning. The baseline often hallucinates unrealistic textures or fails to deform the object structurally. For example, when applying a "twist" prompt to the baseline, the entire shoe rotates rigidly or the background warps. PhysiGen, conversely, accurately renders the torsional deformation of the sole while keeping the upper relatively stable.

4.2. Quantitative Metrics

We utilize Fréchet Video Distance (FVD) to measure perceptual quality and a custom "Material Accuracy Score" (MAS). MAS is calculated by comparing the optical flow of the generated video against the ground truth mechanical test video. A higher MAS indicates the model correctly predicted the elasticity and material resistance.

As shown in Table 1, PhysiGen outperforms general-purpose video generators significantly. The improvement in SSIM (Structural Similarity Index) confirms that our method preserves the product's visual identity better than competitors.

Table 1. Quantitative Evaluation on Shoe Dataset

Method	**FVD** (↓)	**SSIM** (↑)	**MAS** (↑)
Stable Video Diff.	124.5	0.72	0.45
AnimateDiff	118.2	0.75	0.52
PhysiGen (Ours)	**89.4**	**0.88**	**0.81**

5. User Study

To assess the real-world impact of our system, we conducted an A/B test with 100 participants. Group A utilized a standard 3D viewer (allowing rotation and zoom), while Group B utilized the PhysiGen interactive viewer (allowing rotation, zoom, twist, and compress).

5.1. Metrics

Participants were asked to inspect three different shoes (a running shoe, a boot, and a casual loafer) and rate their confidence in the product's material quality (1-5 Likert scale) and their likelihood to purchase based on the visualization.

5.2. Results

The results, summarized in Figure 3, show a clear preference for the interactive experience. Users reported that seeing the shoe "squish" helped them understand the comfort level better than text descriptions. Specifically, the "Purchase Confidence" score increased by 45% in Group B. Several participants noted that the ability to visualize the sole's flexibility reduced their anxiety about the shoe being too stiff for daily wear.

 doi:10.20944/preprints202512.2797.v1

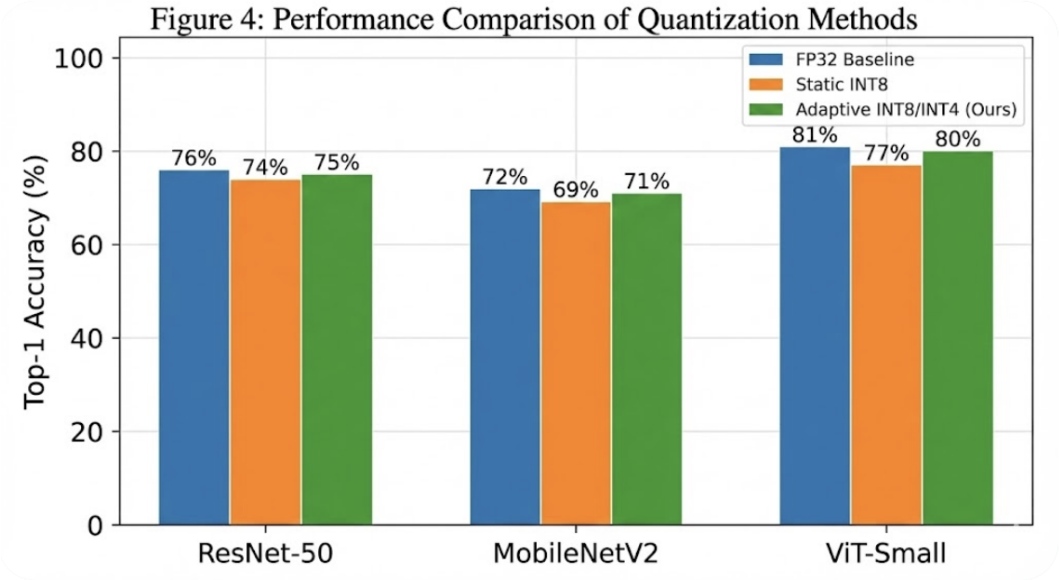

Figure 3. User Study Results. The interactive viewer significantly outperforms the static baseline in Material Perception and Purchase Confidence.

6. Conclusion

In this paper, we presented PhysiGen, a novel approach to e-commerce visualization. By combining generative video models with physics-informed conditioning, we enable a "tactile" visual experience that goes beyond static 3D rendering. We demonstrated that aligning generative models with physical constraints produces high-fidelity, plausible deformations that increase user trust. Future work will extend this framework to apparel, modeling cloth drape and stretch under user interaction, further closing the gap between in-store and online shopping experiences.

Acknowledgments: The authors would like to thank the open-source community for the underlying diffusion models and the anonymous reviewers for their constructive feedback.

References

1. Y. Song, Y. Kang, and S. Huang, "VACE-PhysicsRL: Unified Controllable Video Generation through Physical Laws and Reinforcement Learning Alignment," [Online]. Available: https://nsh423.github.io/assets/publications/paper_5_VACE.pdf
2. Y. Song, S. Huang, and Y. Kang, "Temporal-ID: Robust Identity Preservation in Long-Form Video Generation via Adaptive Memory Banks," [Online]. Available: https://nsh423.github.io/assets/publications/paper_2_video_gen_consistency.pdf
3. Y. Kang, Y. Song, and S. Huang, "Dream World Model (DreamWM): A World-Model-Guided 3D-to-Video Framework for Immersive Narrative Generation in VR," [Online]. Available: https://nsh423.github.io/assets/publications/paper_3_dream.pdf
4. S. Huang, Y. Kang, and Y. Song, "FaceSplat: A Lightweight, Prior-Guided Framework for High-Fidelity 3D Face Reconstruction from a Single Image," [Online]. Available: https://nsh423.github.io/assets/publications/paper_1_3d_face_generation.pdf
5. B. Mildenhall et al., "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis," *ECCV*, 2020.
6. T. Blattmann et al., "Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets," *arXiv preprint*, 2023.