Counterfactual Simulation-Guided Physical Reasoning for Physics-Based Video Generation

Sohyeong Kim* Woo Hyeong Kim* In-Hwan Jin* Hyeongju Mun Kyeongbo Kong†
Pusan National University

Project Page: https://annon-repo-for-science.github.io/CounterFactual/

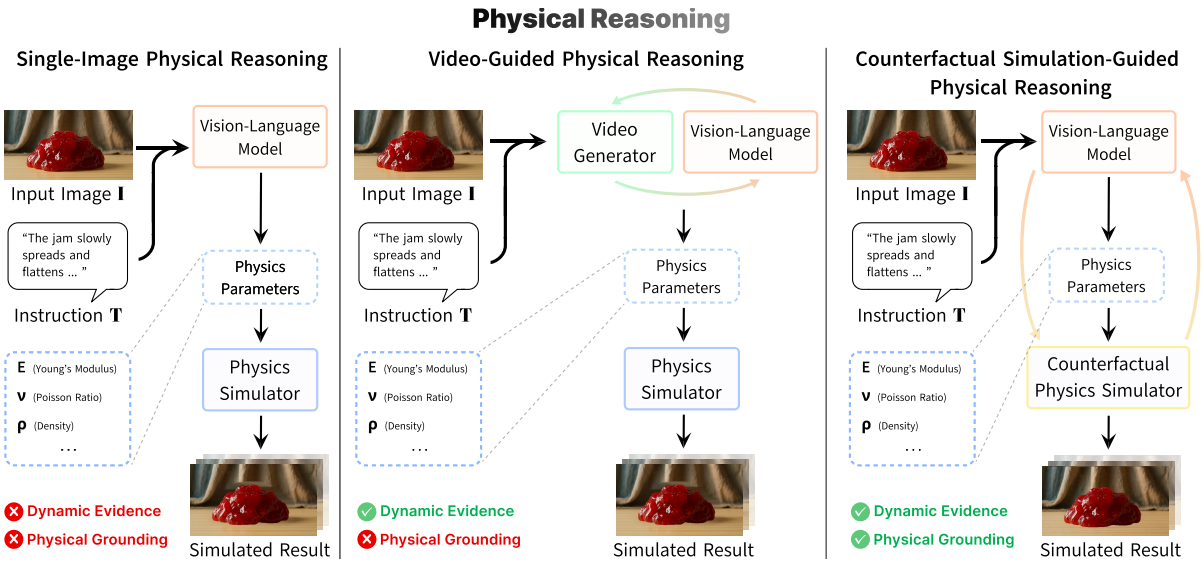

Figure 1. **Comparison of physical reasoning paradigms.** Single-image reasoning relies on static visual evidence, while video-guided reasoning incorporates generated motion as additional dynamic evidence. Our counterfactual-simulator-guided reasoning instead uses controlled physical interventions to provide physically grounded, parameter-specific evidence.

Abstract

Physics-based video generation combines physics simulators with generative models to produce visually realistic, physically grounded dynamics. However, the resulting motion critically depends on simulator parameters, whose estimation from a single image remains inherently ambiguous. We introduce counterfactual physical reasoning, which reasons about physical parameters through controlled interventions and their resulting dynamics. Across 30 scenes, our module improves instruction alignment, visual quality, and physical realism over both baselines, and receives 80.0–85.1% preference in a 50-participant 2AFC study.

1. Introduction

Recent advances in video generation [1–8] and 4D scene generation [9–16] enable highly realistic dynamic visual content, but provide limited guarantees of physical consistency. To incorporate explicit physics, recent hybrid frameworks [17–19] integrate physics simulators with generative models and use simulated dynamics to guide generation. The quality of this physical guidance, however, depends critically on simulator parameters such as friction, restitution, and elasticity, making appropriate parameter estimation essential for reliable physics-based generation.

Determining appropriate simulator parameters traditionally relies on manual specification and iterative trial-and-error. Recent approaches instead leverage VLM-based physical reasoning to infer physical properties from visual observations. As illustrated in Fig. 1, these approaches primarily differ in the evidence available for physical reasoning. *Single-image physical reasoning* infers physical parameters from static visual evidence [19], but appearance alone may be insufficient to determine the intended physical behavior. *Video-guided physical reasoning* further incorporates synthesized motion as dynamic evidence [20], reducing the ambiguity of static observations. However, such generated evidence lacks explicit physical grounding

*Equal contribution. †Corresponding author.

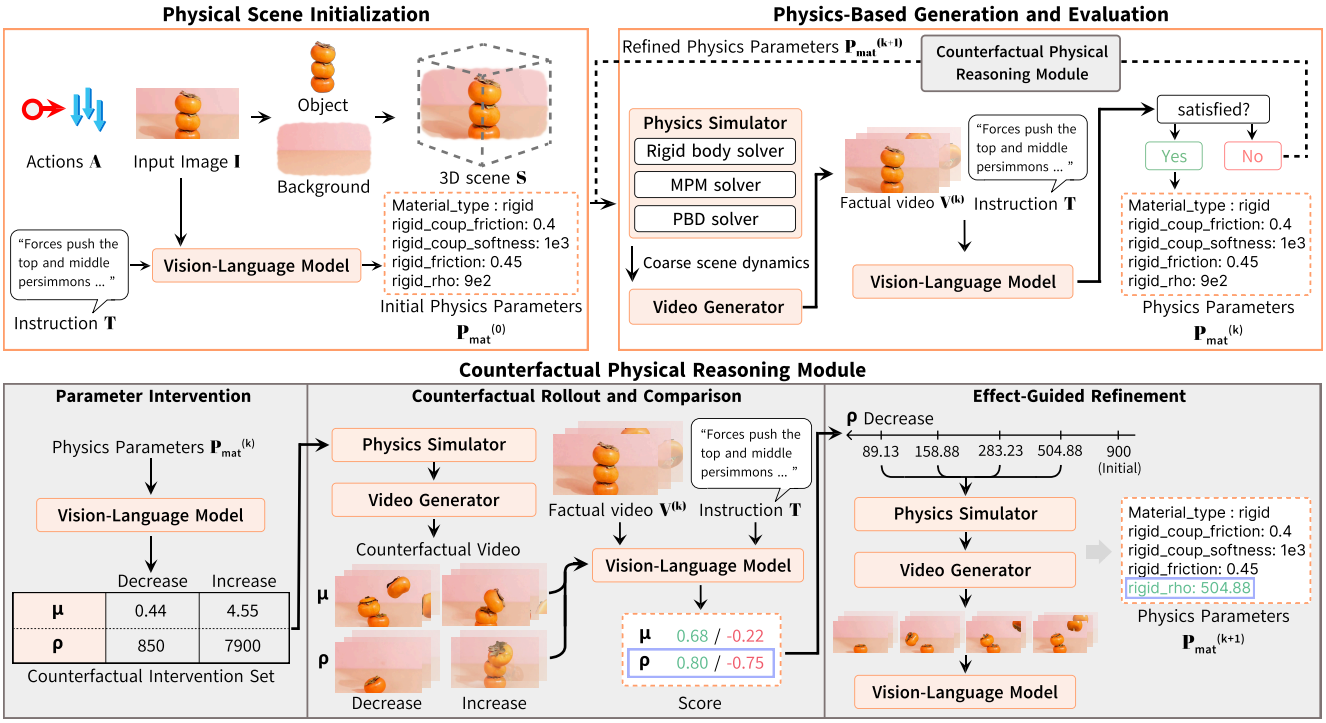

Figure 2. **Overview of our Counterfactual Physical Reasoning framework.** When the current dynamics do not satisfy the user instruction, we construct parameter-wise interventions and simulate their counterfactual outcomes. Comparing the factual and counterfactual dynamics provides physically grounded, parameter-specific evidence for effect-guided refinement toward the intended behavior.

and may not faithfully reflect the effects of the underlying physical parameters.

To address this limitation, we introduce *counterfactual physical reasoning* for physics-based video generation. Counterfactual reasoning examines how an observed outcome would change under an intervention while holding the remaining conditions fixed [21, 22]. We apply this principle by changing one physical parameter at a time and observing the resulting dynamics through physics simulation. Unlike generated motion evidence, these controlled interventions explicitly associate each parameter change with its resulting physical response, providing physically grounded, parameter-specific evidence for physical reasoning.

Based on this principle, we design a *Counterfactual Physical Reasoning Module* that compares parameter-wise counterfactual responses and uses their relative effects to guide physical parameter refinement toward the user-specified behavior.

2. Method

2.1. Overview

Given an input image I, a user instruction T, and an action sequence A, our goal is to generate a physically plausible video whose dynamics align with the user-specified behavior. We denote the object-level material parameters as

$$\mathbf{P}_{\text{mat}} = \{\rho, E, \nu, \sigma_y, \ldots\}, \tag{1}$$

where ρ, E, ν, and σ_y denote density, Young's modulus, Poisson's ratio, and yield stress, respectively.

Our framework first constructs a simulatable 3D scene and initializes its material parameters (Sec. 2.2). The current material configuration is then executed through physics-based generation, and a VLM evaluates the resulting dynamics against the user instruction (Sec. 2.3). When further refinement is required, our **Counterfactual Physical Reasoning Module** (Sec. 2.4) evaluates parameter-wise counterfactual responses and refines the material configuration toward the intended physical behavior. The refined configuration is then used to generate new dynamics, which are evaluated again by the VLM until the intended behavior is achieved.

2.2. Physical Scene Initialization

Following RealWonder [19], we reconstruct a simulatable 3D scene $\mathcal{S} = \mathcal{B} \cup \mathcal{O}$ from the input image I, where $\mathcal{B}$ and $\mathcal{O}$ denote the static background and dynamic objects, respectively. The background provides static collision boundaries, while the objects are represented for subsequent physics simulation.

In parallel, a pretrained VLM Φ_{VLM} estimates the initial object-level material parameters from the input image and user instruction:

$$\mathbf{P}^{(0)}_{\text{mat}} = \Phi_{\text{VLM}}(I, T). \tag{2}$$

The reconstructed scene $\mathcal{S}$ and $\mathbf{P}_{\text{mat}}^{(0)}$ define the initial physical state for simulation.

2.3. Physics-Based Generation and Evaluation

At each iteration k, the current material configuration $\mathbf{P}_{\text{mat}}^{(k)}$ is assigned to the reconstructed scene $\mathcal{S}$ and executed by the physics simulator under the action sequence A. The simulator produces coarse scene dynamics, from which an RGB preview and optical flow are rendered and provided as physical guidance to the video generator. We denote this complete simulation and generation process by $\mathcal{G}$:

$$V^{(k)} = \mathcal{G}\left(\mathcal{S}, A; \mathbf{P}_{\text{mat}}^{(k)}\right). \tag{3}$$

A VLM evaluator Φ_{eval} compares the generated dynamics $V^{(k)}$ with the user instruction T and determines whether the intended behavior is sufficiently satisfied. If satisfied, the current video is returned as the final output. Otherwise, we invoke the Counterfactual Physical Reasoning Module (Sec. 2.4) to further reason about the current physical configuration.

2.4. Counterfactual Physical Reasoning Module

When further refinement is required, reasoning from the current dynamics alone does not reveal how individual material parameters affect the observed physical behavior. Inspired by the intervention principle of counterfactual reasoning [21, 22], our module evaluates alternative physical responses by changing one material parameter at a time while keeping the remaining conditions fixed. It consists of three steps: parameter intervention, counterfactual rollout and comparison, and effect-guided refinement.

Parameter Intervention. For each material parameter, we construct two alternative configurations by decreasing or increasing its current value while keeping all other parameters unchanged. A VLM determines a parameter-specific intervention magnitude $\delta_i^{(k)}$ based on the current configuration and the valid range of each physical parameter. For example, for Young's modulus E, we construct

$$E^{-} = E^{(k)} - \delta_E^{(k)}, \qquad E^{+} = E^{(k)} + \delta_E^{(k)}. \tag{4}$$

The same procedure is independently applied to each parameter in $\mathbf{P}_{\text{mat}}^{(k)}$. We collectively refer to these parameter-wise alternatives as the *counterfactual intervention set*.

Counterfactual Rollout and Comparison. Each configuration in the counterfactual intervention set is executed through the same physics-based generation process $\mathcal{G}$. We treat the current video $V^{(k)}$ as the factual outcome and the videos generated from the interventions as counterfactual outcomes. A counterfactual evaluator Φ_{CF} compares each counterfactual outcome with the factual outcome under the user instruction T and assigns a relative improvement score to each intervention. Positive scores indicate improvement over the factual outcome, while negative scores indicate degradation. These scores indicate both which material parameter most affects the intended behavior and whether increasing or decreasing it is favorable.

Effect-Guided Refinement. We select the material parameter and increase/decrease direction with the strongest counterfactual improvement. After this coarse parameter-wise reasoning, we perform a fine-grained search over the selected parameter. Specifically, candidate values are sampled between its current value and the corresponding parameter bound in the selected direction, using either linear or logarithmic spacing depending on the parameter range. Each candidate is evaluated through the same physics-based generation process, and the VLM selects the candidate whose resulting dynamics best match the user instruction. This yields the refined configuration $\mathbf{P}_{\text{mat}}^{(k+1)}$, while keeping the remaining parameters fixed. The refined configuration is then re-evaluated in the next iteration.

Table 1. Quantitative comparison with physics-based video generation baselines. Higher values are better for all metrics.

Method	CLIP Sim. ↑	Aesthetic ↑	PhysReal ↑
RealWonder [19]	0.269	53.60	0.624
RealWonder + PhyMAGIC [20]	0.273	54.28	0.635
RealWonder + Ours	**0.278**	**55.14**	**0.647**

3. Experiments

Experimental Setup. We evaluate our method on 30 scenes from evaluation sets used in prior physics-based action-conditioned generation works [17–19], covering diverse materials and physical interactions. We compare against RealWonder [19] and RealWonder+PhyMAGIC, where the physical reasoning strategy of PhyMAGIC [20] is incorporated into the RealWonder parameter estimation stage. All methods share the same input image, instruction, action sequence, and physics-based generation pipeline.

Implementation Details. We use GPT-5.6 Luna with high reasoning effort for all VLM components (Φ_{VLM}, Φ_{eval}, and Φ_{CF}), with task-specific prompts for each stage. For effect-guided refinement, we evaluate five candidate values using parameter-specific linear or logarithmic spacing. We perform at most three iterations with early termination when the generated dynamics satisfy the user instruction.

3.1. Quantitative Comparison

We evaluate generated videos from three perspectives: instruction alignment, visual quality, and physical realism. We report CLIP Similarity for video–instruction alignment, Aesthetic Score for perceptual quality, and PhysReal [23] for physical realism, using a GPT-5.5-based evaluator for PhysReal.

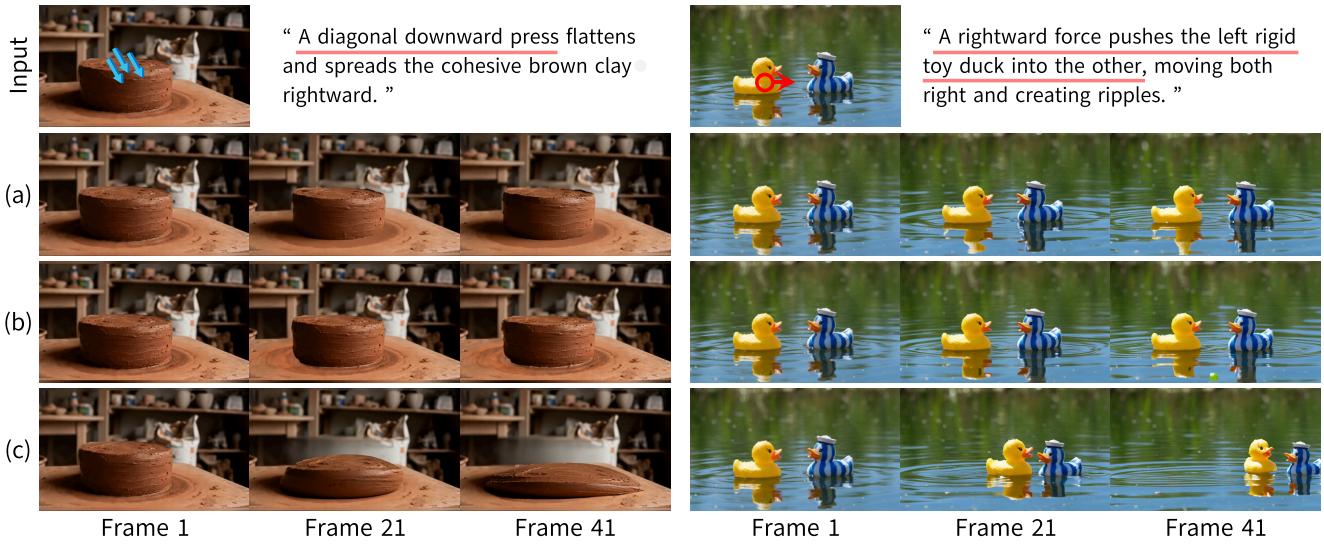

Figure 3. **Qualitative comparison with physics-based video generation baselines.** Given the same input image, instruction, and action sequence, we compare (a) RealWonder [19], (b) RealWonder + PhyMAGIC [20], and (c) RealWonder+ Ours. Our counterfactual physical reasoning produces more instruction-aligned physical responses while preserving visual quality.

Table 2. **2AFC user study with 50 participants.** We report the percentage of participants who prefer our method over each baseline in terms of physical plausibility and instruction alignment.

Comparison	Physical Plaus. ↑	Instr. Align. ↑
Ours vs. RealWonder [19]	81.0%	85.1%
Ours vs. RealWonder + PhyMAGIC [20]	80.0%	83.8%

As shown in Table 1, RealWonder+PhyMAGIC consistently improves over RealWonder, demonstrating the benefit of incorporating dynamic evidence into physical reasoning. RealWonder+Ours achieves the best overall performance, with clear gains in instruction alignment and physical realism while preserving visual quality. These results suggest that counterfactual physical reasoning provides more reliable, physically grounded evidence for determining appropriate physical parameters than reasoning solely from static or generated visual observations.

3.2. User Study

We further conduct a 2AFC user study with 50 participants. For each comparison, participants view two anonymized videos generated from the same input image, instruction, and action, with randomized presentation order, and select the video with (1) more physically plausible dynamics and (2) better instruction alignment.

As shown in Table 2, our method is strongly preferred over both baselines on both criteria. The consistent preference over both RealWonder and RealWonder+PhyMAGIC indicates that the benefits of counterfactual physical reasoning are also reflected in human judgments of physical plausibility and instruction alignment.

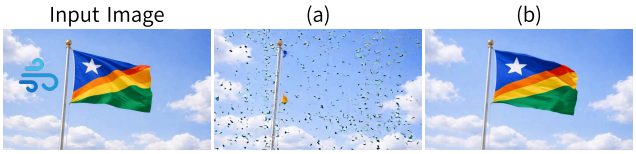

Figure 4. **Ablation of the Counterfactual Physical Reasoning Module.** (a) Ours without the Counterfactual Physical Reasoning Module and (b) Ours. Removing the module leads to unstable dynamics, while the full model produces coherent, instruction-aligned physical behavior.

3.3. Qualitative Comparison

As shown in Fig. 3, RealWonder and RealWonder+PhyMAGIC often fail to sufficiently realize the physical behavior specified by the instruction. For the clay example, both baselines preserve much of the original shape, whereas Ours produces the pronounced flattening and rightward spreading requested by the instruction. Similarly, in the duck interaction, Ours produces a clearer rightward displacement following the applied force. Overall, our method produces physical responses that more faithfully reflect the intended behavior.

3.4. Ablation Study

We ablate the Counterfactual Physical Reasoning Module by allowing the VLM to directly update physical parameters from the current dynamics. As shown in Fig. 4, simulation alone provides physical grounding but does not reveal the effects of individual parameters. Our counterfactual reasoning instead isolates parameter-wise effects, enabling more reliable physical reasoning and instruction-aligned dynamics.

References

[1] Jason Baldridge, Jakob Bauer, Mukul Bhutani, Nicole Brichtova, Andrew Bunner, Lluis Castrejon, Kelvin Chan, Yichang Chen, Sander Dieleman, Yuqing Du, et al. Imagen 3. *arXiv preprint arXiv:2408.07009*, 2024. 1

[2] Midjourney. Midjourney. https://www.midjourney.com/, 2023.

[3] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. *OpenAI Technical Report*, 2024.

[4] 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.

[5] Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao Chen. Gamegen-x: Interactive open-world game video generation. In *International Conference on Learning Representations*, volume 2025, pages 37546–37593, 2025.

[6] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pages 83048–83077, 2025.

[7] Zhenghao Zhang, Junchao Liao, Menghao Li, ZuoZhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2063–2073, 2025.

[8] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 conference papers*, pages 1–11, 2024. 1

[9] Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, et al. Text-to-4d dynamic scene generation. *arXiv preprint arXiv:2301.11280*, 2023. 1

[10] Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B Lindell. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7996–8006, 2024.

[11] Yufeng Zheng, Xueting Li, Koki Nagano, Sifei Liu, Otmar Hilliges, and Shalini De Mello. A unified approach for text- and image-guided 4d scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7300–7309, 2024.

[12] Sherwin Bahmani, Xian Liu, Wang Yifan, Ivan Skorokhodov, Victor Rong, Ziwei Liu, Xihui Liu, Jeong Joon Park, Sergey Tulyakov, Gordon Wetzstein, et al. Tc4d: Trajectory-conditioned text-to-4d generation. In *European Conference on Computer Vision*, pages 53–72. Springer, 2024.

[13] Bohan Zeng, Ling Yang, Siyu Li, Jiaming Liu, Zixiang Zhang, Juanxi Tian, Kaixin Zhu, Yongzhen Guo, Fu-Yun Wang, Minkai Xu, et al. Trans4d: Realistic geometry-aware transition for compositional text-to-4d synthesis. *arXiv preprint arXiv:2410.07155*, 2024.

[14] Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T. Barron, and Aleksander Holynski. Cat4d: Create anything in 4d with multi-view video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26057–26068, 2025.

[15] In-Hwan Jin, Haesoo Choo, Seong-Hun Jeong, Park Heemoon, Junghwan Kim, Oh-joon Kwon, and Kyeongbo Kong. Optimizing 4d gaussians for dynamic scene video from single landscape images. In *The Thirteenth International Conference on Learning Representations*, 2025.

[16] Hyeongju Mun, In-Hwan Jin, Sohyeong Kim, and Kyeongbo Kong. Livingworld: Interactive 4d world generation with environmental dynamics. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026. 1

[17] Zizhang Li, Hong-Xing Yu, Wei Liu, Yin Yang, Charles Herrmann, Gordon Wetzstein, and Jiajun Wu. Wonderplay: Dynamic 3d scene generation from a single image and actions. *arXiv preprint arXiv:2505.18151*, 2025. 1, 3

[18] Jiahao Zhan, Zizhang Li, Hong-Xing Yu, and Jiajun Wu. Perpetualwonder: Long-horizon action-conditioned 4d scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

[19] Wei Liu, Ziyu Chen, Zizhang Li, Yue Wang, Hong-Xing Yu, and Jiajun Wu. Realwonder: Real-time physical action-conditioned video generation. *arXiv preprint arXiv:2603.05449*, 2026. 1, 2, 3, 4

[20] Siwei Meng, Yawei Luo, and Ping Liu. Phymagic: Physical motion-aware generative inference with confidence-guided llm. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026. 1, 3, 4

[21] Judea Pearl. *Causality*. Cambridge university press, 2009. 2, 3

[22] Judea Pearl. The causal foundations of structural equation modeling. *Handbook of structural equation modeling*, pages 68–91, 2012. 2, 3

[23] Boyuan Chen, Hanxiao Jiang, Shaowei Liu, Saurabh Gupta, Yunzhu Li, Hao Zhao, and Shenlong Wang. Physgen3d: Crafting a miniature interactive world from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6178–6189, 2025. 3