PerpetualWonder: Long-Horizon Action-Conditioned 4D Scene Generation

Jiahao Zhan*,† Zizhang Li* Hong-Xing Yu Jiajun Wu

Stanford University

https://johnzhan2023.github.io/PerpetualWonder/

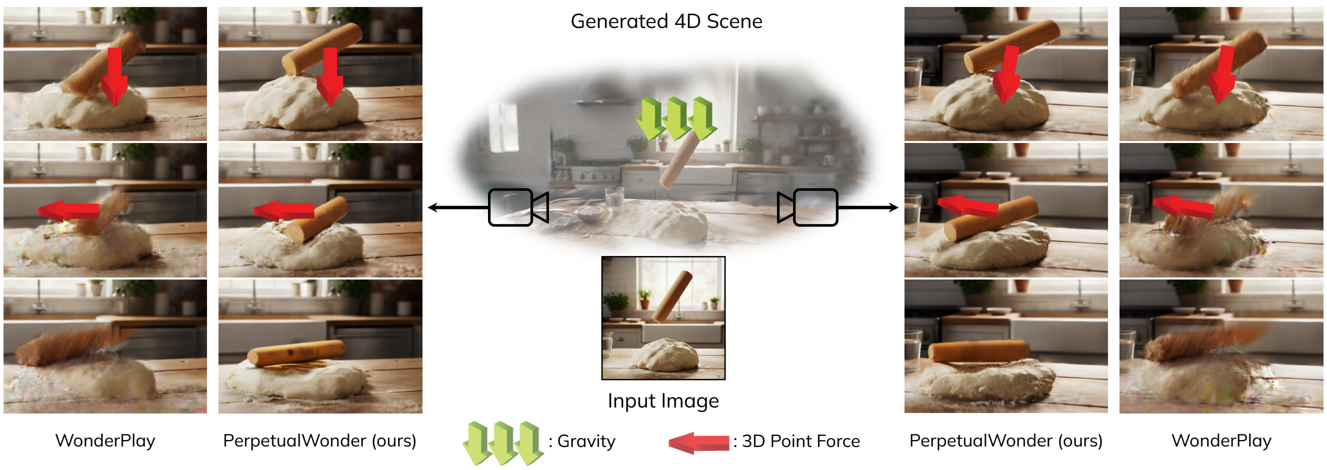

Figure 1. We propose **PerpetualWonder**, a hybrid generative simulator that generates a 4D scene with long-horizon actions and a single image. Here we show a side-by-side comparison for a three-step action sequence (top to bottom, actions overlaid on the images). The left and right image blocks show renderings from two different viewpoints. PerpetualWonder shows superior performance over the previous method. We show video results in https://johnzhan2023.github.io/PerpetualWonder/.

Abstract

We introduce PerpetualWonder, a hybrid generative simulator that enables long-horizon, action-conditioned 4D scene generation from a single image. Current works fail at this task because their physical state is decoupled from their visual representation, which prevents generative refinements to update the underlying physics for subsequent interactions. PerpetualWonder solves this by introducing the first true closed-loop system. It features a novel unified representation that creates a bidirectional link between the physical state and visual primitives, allowing generative refinements to correct both the dynamics and appearance. It also introduces a robust update mechanism that gathers supervision from multiple viewpoints to resolve optimization ambiguity. Experiments demonstrate that from a single image, PerpetualWonder can successfully simulate complex, multi-step interactions from long-horizon actions, maintaining physical plausibility and visual consistency.

*Equal contribution. †Work was done when J. Zhan was a visiting student at Stanford University. J. Zhan is currently with Fudan University.

1. Introduction

Recent years have seen remarkable progress in generative models for text, images, and videos. This rapid advancement motivates the creation of generative world models [4, 15, 39], which are crucial for applications in VR/AR, gaming, and embodied AI [30]. A key capability for such models is not just to generate realistic content, but to simulate a world that responds to user actions. We study the task of **action-conditioned 4D scene generation from a single image**. Given a single input image and a sequence of physical actions (local forces like pushes and pokes, or global forces like wind fields and gravity), the goal is to generate the dynamic 4D scene that corresponds to the actions and evolves plausibly over time.

Early attempts to generate 4D content in response to actions relied heavily on traditional physical simulation [6, 14, 48, 58]. These methods, while offering precise and interpretable physical control, are driven entirely by the traditional simulator for both dynamics and appearance. This often results in a significant realism gap, as simplified physics and analytic rendering struggle to capture the complex visual phenomena of the real world, such as subtle material deformations, lighting changes, and secondary visual effects like splashes. Concurrently, modern video generation models [4, 44] have become incredibly powerful, learning strong priors about real-world dynamics and appearance from massive video data.

This presents a new opportunity, leading to the rise of the **hybrid generative simulator** [21, 38]: a system that first uses the traditional physical simulation to generate coarse, action-conditioned dynamics, and then employs a video generation model as a neural refiner to achieve high-fidelity visual realism. The hybrid generative simulator aims for the best of both worlds, combining the strengths of traditional physical simulators, including consistency and controllability, with the power of video generation models, which provide visual realism and complex dynamics.

Recent WonderPlay [21] is a realization of this hybrid generative simulator concept. However, its approach is fundamentally limited to short-term interactions within a single time window. The core problem is that the flow of information is incomplete: the physical state informs the video model, but the video model's refinement only propagates back to the scene's appearance representation, not its underlying physical state. The physical and visual representations are thus decoupled. This prevents any form of long-horizon, sequential interaction, as the physical simulator is blind to the generative corrections from the previous step, leading to the accumulation of errors.

We aim to overcome this fundamental limitation and enable long-horizon, sequential actions. This requires a system that can perpetually cycle between user actions, physical simulation, and generative refinement. We identify two fundamental challenges: the first is that current physical states (physics particle position and velocity) cannot be updated by the refinement from the video generation model. A new representation is required to unify the physical and visual domains. Then, to update the unified representation, the refinement from video generation models must be multi-view to prevent ambiguity in optimization. However, the video models naturally will not generate perfectly consistent videos from different viewpoints. To resolve this ambiguity, a robust update mechanism is required.

To address these challenges, we propose **PerpetualWonder**, a new hybrid generative simulator for long-horizon action-conditioned 4D scene generation, as shown in Figure 1. First, we introduce *visual-physical aligned particle* (VPP), a novel unified representation that tightly binds physics particles to the visual representations. The proposed VPP acts as a bidirectional bridge: enabling the forward physics pass that uses physical simulation to drive the visual prediction, and critically, updating the physics particles through the optimized visual representation in a backward optimization, resulting in an innovative closed-loop system. Next, we propose a multi-view optimization mechanism to ensure the update is 3D consistent and plausible. We first initialize a complete 3D scene from the input image using dense view generation. This initialization allows us to render the scene from arbitrary viewpoints in a wide range and use the video model to gather supervision from multiple views. Then we progressively leverage refined videos from multiple viewpoints for backward optimization. This strategy resolves ambiguity, producing a 4D scene that is both visually realistic and physically coherent, ready for the next user action. In summary, our contributions are:

- We tackle the task of long-horizon action-conditioned 4D scene generation, enabling sequential action interactions.
- We propose PerpetualWonder, a novel hybrid generative simulator that features a unified representation for both physical state and visual appearance, and a multi-view optimization mechanism for consistent scene updates.
- We demonstrate that PerpetualWonder consistently outperforms prior work in action-conditioned 4D scene generation, including both long-horizon interaction abilities and scene consistency.

2. Related Work

Dynamic 4D scene generation. Our work generates action-conditioned 4D scenes and connects to a rich body of research on dynamic scene representation and generation. Early work in this domain focused on reconstructing dynamic scenes from real-world captures. Representations rapidly evolved from dynamic Neural Radiance Fields (NeRF) [23, 27–29] to dynamic Gaussian Splatting [18, 45, 49, 51]. While these methods achieve high-fidelity rendering of complex motion, they are fundamentally limited to replaying pre-captured events and do not support user actions or the simulation of novel dynamics.

More recently, a dominant line of research has focused on generative models for synthesizing novel 4D content. Many of these approaches distill the powerful priors from large-scale video models to generate 4D animations from text or image prompts [1, 2, 9, 22, 32, 37, 56]. These methods leverage dynamic 3D representations to create temporally consistent animations. Other works focus on directly modeling the 4D space-time volume [43, 46, 60] or parameterizing 4D representations with generative networks [33, 57]. However these approaches share a critical limitation: the synthesized dynamics are passive. They generate pre-determined animations and lack the mechanisms to simulate diverse, physically plausible responses to user input actions, which is the core focus of our work.

Physics-grounded scene generation. To enable action-conditioned interactions, a separate line of work has focused on integrating physical principles [20] and traditional physical simulation methods into the scene generation process.

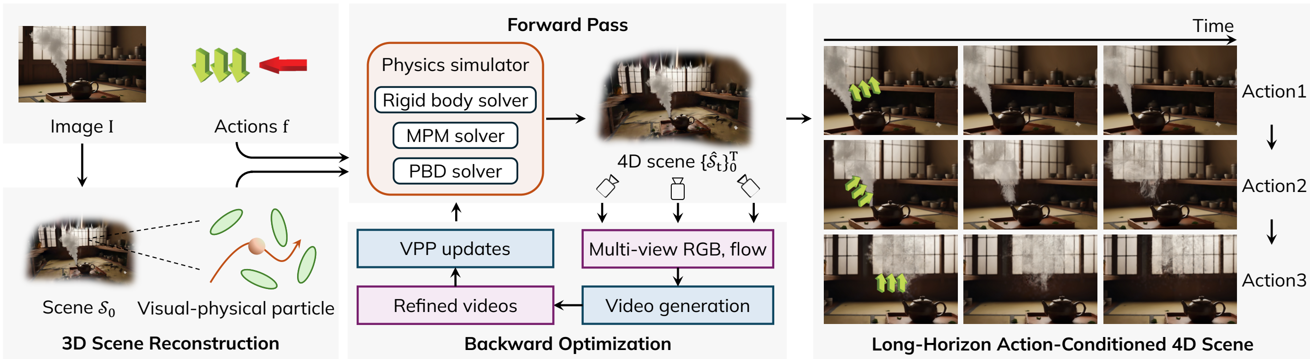

Figure 2. **Overview of PerpetualWonder**. Given an input image, based on the visual-physical aligned particle, we reconstruct a 3D scene from synthesized dense views. Then we iterate between a forward physics pass and a backward neural optimization. The forward pass leverages physical simulation to generate coarse scene dynamics. Then the backward optimization updates the scene according to the multi-view refined videos from the video generation model. The closed-loop system enables long-horizon actions for the final 4D scene generation. The rendered results on the right showcase the generated scene from each consecutive action.

Early methods relied entirely on traditional physical simulation [6, 14, 48, 58], which provides precise, interpretable control but suffers from a significant realism gap. These simulators often use simplified, approximated physics and rendering with fixed visual appearance, struggling to capture the visual phenomena of the real world, such as subtle material deformations and secondary visual effects.

To bridge this realism gap, recent works have begun to integrate physics with the strong priors of generative models. This has culminated in the hybrid generative simulator [21, 38] approaches, which aim to blend the controllability of the physical simulators with the visual realism of the video generation models. The most relevant work, WonderPlay [21], first uses the physics solvers to generate coarse, action-conditioned dynamics and then employs a video model as a neural refiner to achieve high-fidelity visual realism. However, these methods are limited by a fundamental architectural drawback that the flow of information is incomplete. The generative refinements, whether for appearance or dynamics, only affect the visual primitives and do not propagate back to the underlying physical state. Furthermore, this makes these methods fundamentally limited to short-term interactions, and are unable to support the long-horizon actions that our work aims to solve.

Controllable video generation. Concurrently, video generation models have become incredibly powerful [4, 13, 40, 44, 50, 62], achieving stunning realism, but controlling their output remains a significant challenge. Most existing control methods focus on non-physical aspects, such as following text instructions, camera trajectories [3, 34, 41, 63], or 2D motion guidance through keypoints and trajectories [7, 10, 26, 36, 47, 59]. While some works explore conditioning on 2D force vectors [11] to mimic the real-world actions, these models lack an explicit underlying 3D representation. These 2D-centric video approaches are insufficient for our task, as they cannot ensure physically accurate action conditions within a 3D scene or guarantee 3D consistency when rendering the resulting 4D scenes from novel viewpoints. In contrast, our method operates on a full 3D representation, providing the structure for both precise physical interaction and consistent multi-view rendering.

3. PerpetualWonder

Our goal is to achieve long-horizon action-conditioned 4D scene generation from a single image $\mathbf{I}$. Given a sequence of user actions $\{\mathcal{A}_t\}_{t=0}^{T-1}$ including the global force $\mathbf{f}(x, y, z, t)$ (e.g., gravity, wind field) and/or the local force $\mathbf{f}(t)$, our proposed PerpetualWonder outputs a dynamic 4D scene sequence $\{\mathcal{S}_t\}_{t=0}^{T}$. At any time t, the scene state $\mathcal{S}_t = (\mathcal{B}_t, \mathcal{F}_t)$ is decomposed into the background $\mathcal{B}_t$ and the dynamic, interactable foreground $\mathcal{F}_t$.

As illustrated in Figure 2, PerpetualWonder achieves this by an innovative closed-loop hybrid generative simulator, perpetually iterating between a forward physics pass Φ_p and a backward neural optimization pass Ψ_n. To enable this, we must solve the two fundamental challenges: first, crafting a unified representation that allows the physical state to be updated by generative refinements from video models; and second, developing a robust and consistent optimization mechanism to perform these updates without ambiguity.

Accordingly, we first introduce the proposed *visual-physical aligned particle* (VPP) as a unified representation (Section 3.1). We then detail our multi-view optimization mechanism 3.2, which leverages a single image-based dense 3D reconstruction to perform a progressive, multi-view backward optimization. Finally, we assemble these components into the complete PerpetualWonder simulation loop for long-horizon actions (Section 3.3).

3.1. Visual-Physical Aligned Particle

The core of long-horizon action interactions lies in the underlying representation. Previous hybrid generative simulators use decoupled visual primitives (e.g., gaussian splatting [18]) for appearance and physics particles for dynamics. The incomplete binding uses physics particles to drive visual primitives, making it impossible for visual refinements from the video generation model to correct the underlying dynamics. This prevents a closed-loop system and makes long-horizon simulation inapplicable.

To solve this, we introduce the VPP, a novel unified representation that tightly binds the physics particles to the visual primitives, creating a bidirectional bridge between dynamics and appearance.

We define the foreground $\mathcal{F}_t$ as a set of O objects, $\mathcal{F}_t = \{\mathcal{P}_t^o, \mathcal{V}_t^o, \mathcal{G}_t^o\}_{o=1}^O$. For simplicity, here we omit the object index o and time index t. The VPP for a single object consists of:

Physics dynamics. A set of J physics particle positions $\mathcal{P} = \{p_j\}_{j=1}^J$ and their velocities $\mathcal{V} = \{v_j\}_{j=1}^J$. The particles $\mathcal{P}$ are sampled from an initial object mesh volume. We detail the process for obtaining the object mesh in the following 3D scene initialization subsection.

Visual appearance. A set of $J \times K$ visual primitives, i.e., gaussians [18], $\mathcal{G} = \bigcup_{j=1}^J \{g_{j,k}\}_{k=1}^K$. Specifically, each physics particle p_j serves as an anchor to a small set of K gaussian primitives.

The key to the VPP is how these K gaussians are parameterized relative to their anchor particle p_j, enabling both adherence to physics and optimization of appearance and dynamics. We define the attributes of these gaussians with the following details:

- **Position offset** $\tilde{p}$: Each gaussian's 3D position $\mu_{j,k}$ is defined by a small, learnable position offset $\tilde{p}_{j,k}$ from its corresponding physics particle p_j:

$$\mu_{j,k} = p_j + \tanh(\tilde{p}_{j,k}) \cdot \delta, \tag{1}$$

  where δ is the physics particle size defined during the simulator's sampling process.
- **Scale**: Each gaussian primitive is defined to be isotropic, with a fixed scaling value not larger than δ.
- Spatial opacity o_s and temporal opacity $o_t(t)$: Inspired by [42], we define two opacity parameters. o_s is the standard learnable spatial opacity scalar [18]. $o_t(t)$ is a learnable temporal opacity, parameterized by a center time μ_t and duration s_d:

$$o_t(t) = \exp\left(-\frac{1}{2} \times \left(\frac{t - \mu_t}{s_d}\right)^2\right). \tag{2}$$

  The final opacity is $o(t) = o_s \times o_t(t)$.

Rotation and color attributes are parameterized the same as in the original 3D Gaussian Splatting work [18]. This VPP representation forms the foundation of our closed-loop system. By ensuring every physics particle p_j has a corresponding set of visual primitives $\{g_{j,k}\}$, all dynamics and appearance are now expressed by the optimizable visual primitives. This structure creates the bidirectional bridge, where the forward physics pass drives the dynamics by updating p_j, which in turn moves all anchored visual primitives. Critically, the backward optimization pass can now refine the final 4D scene by optimizing the attributes of $\{g_{j,k}\}$ while being constrained to remain consistent with the anchoring physics particles.

3.2. Multi-View Optimization

The proposed VPP serves as a representation that can be updated for both 4D scene dynamics and appearance. The remaining challenge lies in how to consistently perform this update. Simply optimizing from a single-view video refinement, as done in WonderPlay [21], leads to significant ambiguity and artifacts from novel viewpoints. To tackle this, we introduce a robust multi-view optimization mechanism. This mechanism consists of two components. First, we initialize the complete 3D scene from the single input image to enable rendering from arbitrary viewpoints. Second, we leverage this 3D scene to perform a progressive optimization using supervision gathered from multiple views.

3D scene initialization. As the initialization for our hybrid generative simulator, we first reconstruct the 3D scene $\mathcal{S}_0$ from the single input image I. To construct a 3D scene that supports rendering from arbitrary viewpoints, we employ a state-of-the-art camera-controlled video model GEN3C [34] to synthesize dense surrounding views of the scene from the input image. This video is then processed with COLMAP [35] to acquire a scene point cloud for initializing 3D Gaussians.

Following standard 3DGS optimization [18], we obtain the set of N gaussian primitives $\{G_i\}_{i=1}^N$ as the initial scene representation. Each gaussian G_i is parameterized by position p_i, orientation q_i, scale s_i, opacity o_i, and color c_i.

To decompose the scene into background and foreground objects, inspired by Gaussian Grouping [52], we add a learnable feature g_i to each primitive. We leverage SAM2 [31] to obtain object masks on the dense surrounding views, which supervise these learnable features. After decomposition, the Gaussian primitives are split into a background set $\mathcal{B}_0$ and sets for each foreground object. These foreground object gaussians are then transformed into closed-surface meshes using TSDFusion [55]. Then these meshes are used to sample the initial physics particles $\mathcal{P}_0$ for the proposed VPP. This is followed by another round of optimization for the gaussian primitives $\{\mathcal{G}_0^o\}_{o=1}^O$, with respect to the frames from GEN3C [34], and replace the original foreground objects' gaussians.

This approach contrasts with the scene initialization in WonderPlay, which relies on single-view depthunprojection [53, 54] and object placement. Our multi-view reconstruction process builds the background and all objects together in a single, unified 3D coordinate space, which is essential for rendering from dense, arbitrary viewpoints.

Progressive multi-view optimization. The refinement process Ψ_n leverages pre-trained video generation models to refine both appearance and dynamics of the underlying 4D scene. The bridge between our VPP representation and the video generation model is the rendered video. Following [21], the coarse 4D scene from physical simulation (detailed in Section 3.3) is rendered into RGB and optical flow frames from a chosen viewpoint. This coarse video is then refined by the video generation model [5, 50] through a bimodal control scheme, resulting in a refined video $\mathbf{V}_t$.

However, refined videos from different viewpoints are inevitably inconsistent and can not be directly used for optimization. To resolve the ambiguity and acquire the desired 4D scene, we introduce a two-part solution.

First, we design a loss function that, combined with our VPP representation, provides a strong consistency prior. For a given time step t, the overall loss function is:

$$\begin{aligned}\mathcal{L} = &\mathcal{L}_{\mathrm{p}}(\mathtt{Render}(\mathcal{B}_t) \odot (1 - \mathbf{M}), \mathbf{V}_t \odot (1 - \mathbf{M})) \\ &+ \mathcal{L}_{\mathrm{p}}(\mathtt{Render}(\mathcal{G}_t), \mathbf{V}_t \odot M) + \lambda_{\mathrm{sim}}\mathcal{L}_{\mathrm{sim}}.\end{aligned} \tag{3}$$

Here, $\mathbf{M}$ is the binary mask for the foreground VPPs, $\mathtt{Render}(\cdot)$ is the gaussian rendering function, and $\mathcal{L}_p$ is the photometric loss (L1 and SSIM). In practice, we also model the background gaussians $\mathcal{B}_t$ with learnable spatial and temporal opacity (Eq. 2) and other attributes (excluding position) to capture secondary visual effects like shadows. For the foreground VPP $\mathcal{G}_t$, we introduce a simulation consistency loss term $\mathcal{L}_{\mathrm{sim}}$:

$$\mathcal{L}_{\mathrm{sim}} = \frac{1}{T \cdot J} \sum_{t=1}^{T} \sum_{j=1}^{J} \left\| p_{j,t} - \frac{1}{K} \sum_{k=1}^{K} \mu_{j,k,t} \right\|_2^2 \tag{4}$$

The $\mathcal{L}_{\mathrm{sim}}$ penalizes the visual primitives $\mu_{j,k,t}$ for deviating from their corresponding physics particle $p_{j,t}$. The VPP representation and the simulation consistency loss act as a strong regularizer, ensuring the optimized visual primitives do not break apart from their physical anchors, which inherently mitigates inconsistency.

Second, to resolve the remaining visual ambiguity from conflicting refinements in a multi-view setting, we introduce a progressive optimization strategy. To leverage inconsistent multi-view videos without introducing artifacts, we first render and refine the video only from the input image's viewpoint and optimize the scene representation with respect to this single video. Then, we render the 4D scene from other viewpoints and use the video model to refine them with a smaller control weight. Finally, we leverage all refined videos from every viewpoint to optimize the scene representation again, yielding a consistent 4D scene.

3.3. Simulation Loop

With the proposed novel VPP representation and multi-view optimization mechanism now defined, we can assemble the PerpetualWonder simulation loop. This loop operates over a time window of T steps and consists of three key stages: a forward pass to generate the entire sequence, a backward optimization pass to refine it, and a loop closure that enables the next round of action-conditioned interaction.

Forward pass. The loop begins with the forward physics pass Φ_p. Given the scene state $\mathcal{S}_0$, the hybrid generative simulator computes the coarse 4D scene for the time window of T steps. This is achieved by applying the physics operator Φ_p step-by-step for each user action $\mathcal{A}_t$ from $t = 0$ to $T-1$: $\hat{\mathcal{S}}_{t+1} = \Phi_p(\hat{\mathcal{S}}_t, \mathcal{A}_t)$. This process generates a coarse sequence $\{\hat{\mathcal{S}}_t\}_{t=0}^T$. We adopt a set of solvers [17, 21, 24, 25] for various materials, including cloth, sand, snow, liquid, smoke, elastic, and rigid bodies. We refer the reader to Li et al. [21] for more details of the physical simulation process. As aforementioned, this forward pass provides controllable dynamics in response to actions but lacks visual realism and may contain physical inaccuracies.

Backward optimization. The coarse sequence $\{\hat{\mathcal{S}}_t\}_{t=0}^T$ is then fed into the refinement process Ψ_n. As introduced in Section 3.2, we apply our progressive multi-view optimization to the coarse 4D scene. This step leverages video model priors to correct the appearance and dynamics across all T steps, resulting in the final refined sequence $\{\mathcal{S}_t\}_{t=0}^T$.

Loop closure and long-horizon actions. The final step closes the loop and enables perpetual interaction, i.e., performing a new round of actions for the following time window. The final refined state of the current time window, $\mathcal{S}_T$, becomes the starting state $\mathcal{S}_0$ for the next T-step simulation. To prepare the system for this new time window, we update the VPP's underlying physics particles using the results of the backward pass. Specifically, we average the positions of the optimized visual primitives $\{g_{j,k}\}$ at time T to update the position p_j of their corresponding physics particle. The velocity is directly inherited from the original velocities at time T, which is applicable because $\mathcal{L}_{\mathrm{sim}}$ limits the position updates of physics particles in a small range. This corrected physics state $\{\mathcal{P}_T, \mathcal{V}_T\}$ becomes the input for the next forward pass sequence, allowing the system to simulate sequential interactions over a long horizon.

4. Experiment

Implementation details. We initialize the 3D scene $\mathcal{S}_0$ by reconstructing it from 242 views generated by the camera-controlled video model GEN3C [34]. We employ Genesis [19] as our physics simulator for different materials. Our

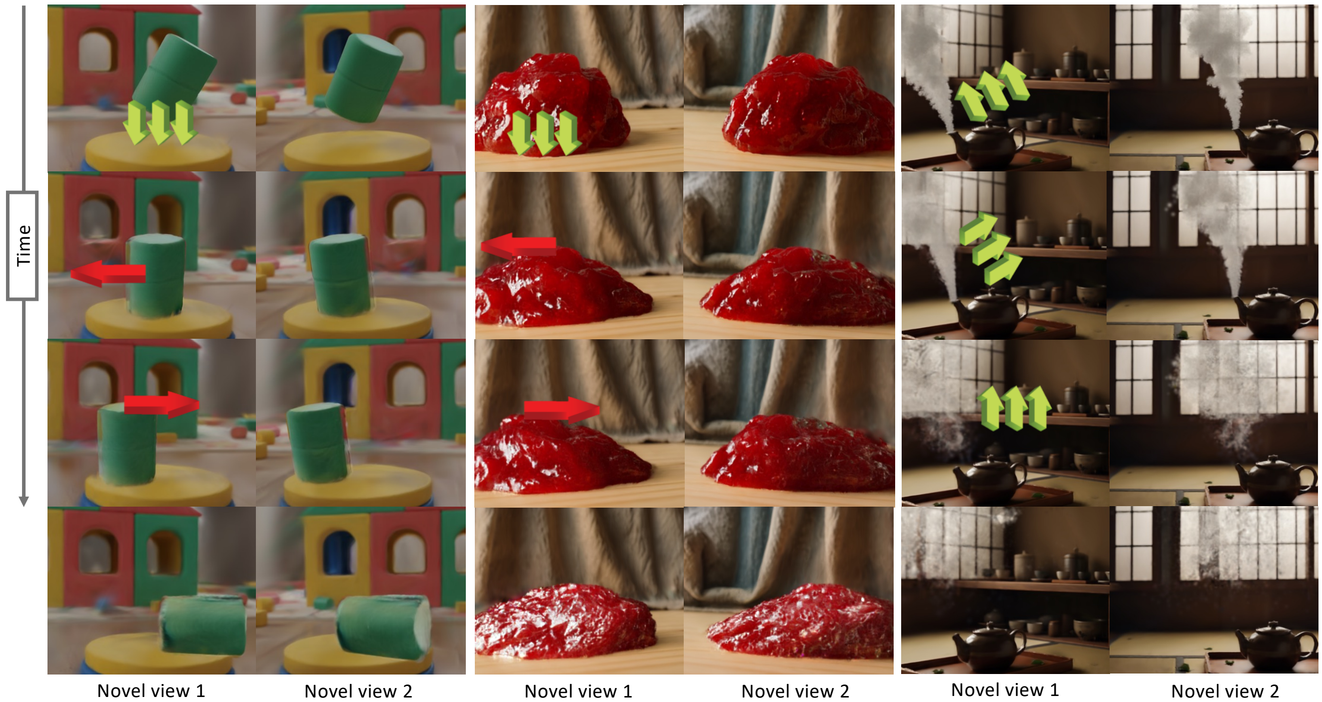

Figure 3. **Qualitative results** of the proposed PerpetualWonder. We show the long-horizon scenes with three consecutive actions. and indicate global force (gravity or wind force field), and 3D point force, respectively. The results are all rendered from novel views, demonstrating our method's ability in long-horizon action-conditioned 4D scene generation.

Method	Camera Ctrl	3D Consist	Imaging
Wan2.2 [40]	59.73	65.35	67.03
GEN3C [34]	80.29	61.69	66.25
WonderPlay [21]	75.95	63.93	36.80
Tora [59]	51.80	60.77	54.37
Wan2.6 [40]	64.75	70.49	66.09
DaS [12]	78.96	62.18	60.23
Veo3.1 [44]	60.61	73.93	**67.82**
PerpetualWonder (ours)	**93.26**	**80.41**	66.98

Table 1. **Quantitative comparison** on 10 scenes using World-Score [8] metrics. Ctrl = Controllability, Consist = Consistency.

	Physics Plausibility	Motion Fidelity
over Wan2.2 [40]	74.1%	71.8%
over GEN3C [34]	93.5%	83.5%
over WonderPlay [21]	80.8%	86.3%
over Veo3.1 [44]	62.0%	70.8%
over Wan2.6 [40]	68.5%	77.3%
over Tora [59]	83.5%	85.3%
over DaS [12]	80.9%	81.9%

Table 2. **2AFC human study results** of favor rate of our PerpetualWonder over baseline methods in dynamic realism.

experimental results typically demonstrate multi-step interactions across 3 time windows, with each window spanning 392 physical simulation steps and receiving different input actions. For refinement, the video generation model [5] is conditioned on RGB and optical flow renderings at the resolution of H=704, W=1280 from the coarse dynamics. The output videos consist of 49 frames, each frame is sampled from every 8 physical simulation steps. The progressive multi-view optimization mechanism uses 3 key views for supervision: the frontal, left-side, and right-side views.

Baselines. We compare PerpetualWonder against two main categories of state-of-the-art methods: conditional video generation models and hybrid generative simulators. For conditional video generators, we evaluate against Wan2.2 [40], Wan2.6, Veo3.1 [44], Tora [59], DaS [12] and GEN3C [34]. For I2V video generators, both the action and camera trajectory are specified via text prompts. For GEN3C, we leverage its native camera control capabilities and embed the desired action within the text prompt. For trajectory-based video models like Tora, the dynamics from physical solvers are utilized to drive the generation. For the hybrid generative simulator, we compare against the most relevant prior work, WonderPlay [21]. To ensure a fair comparison focused on the core simulation loop and representation, we also create a stronger baseline, WonderPlay++. This baseline uses our superior multi-view 3D reconstruction for initialization, which provides a more 3D-consistent scene but retains WonderPlay's original decoupled representation and single-view optimization methodology.

Metrics. Our evaluation is performed on a curated dataset of 10 scenes that feature a diverse range of materials, including cloth, rigid bodies, elastic objects, liquids, gases,

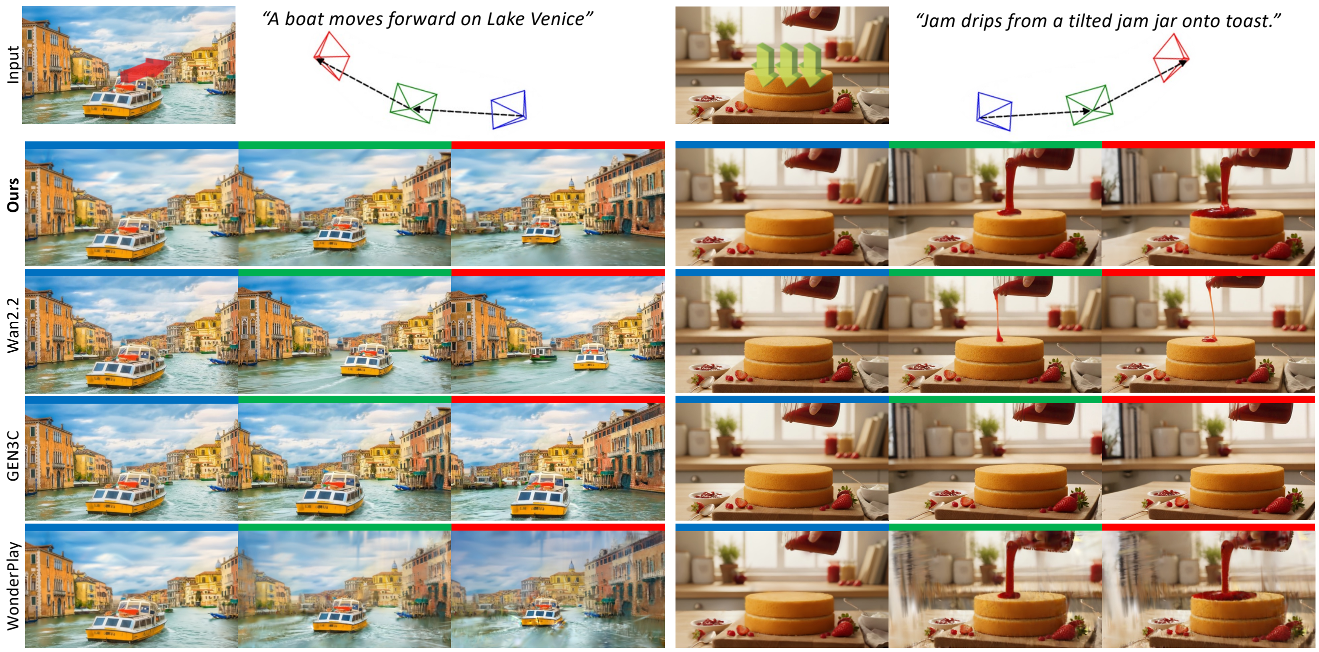

Figure 4. **Qualitative comparisons** between PerpetualWonder (ours) and the baseline methods. The top row shows the input images, actions, camera trajectories, and the texts describing the actions for conditional video generators [34, 40]. For ease of comparison, only one time window is shown. The images from left to right illustrate the resulting scene dynamics and camera motion for each method.

and granular substances. We assess two primary aspects: the quality of the generated 4D scene and the plausibility of the physical dynamics. Please refer to the supplement for the detailed metrics explanations.

4.1. Results

Qualitative comparison to baselines. Figure 4 presents a side-by-side qualitative comparison of our method against some representative baselines. As illustrated, the conditional video generation models fail to support both interaction and arbitrary view changes. Wan2.2 [40] generates plausible motion but disregards camera change instructions provided in the text prompt, causing the camera to remain almost fixed in both scenes. Conversely, GEN3C [34] is 3D-aware and adheres well to camera trajectories, but the object of interest remains static and completely unaffected by the action described in the text prompt. The hybrid generative simulator WonderPlay [21] successfully applies the user actions, but its reliance on single-view optimization leads to severe visual artifacts and geometric inconsistencies when the scene is rendered from novel viewpoints. In contrast, PerpetualWonder generates the complete and consistent 4D scene that both correctly responds to the action inputs and naturally supports rendering from arbitrary viewpoints.

Quantitative results. The quantitative results in Table 1 show PerpetualWonder's significant advantages over both conditional video generators and the hybrid simulator. PerpetualWonder achieves best performance in camera controllability and 3D consistency, while keeping a high level of imaging quality. The user study result is shown in Table 2. In comparison to all baselines, about 70% to 90% of the participants prefer PerpetualWonder across both aspects. This strong preference provides strong evidence that PerpetualWonder successfully generates plausible physical dynamics, enabling long-horizon simulations by preventing the accumulated errors that degrade realism in baseline methods.

Long-horizon actions. A significant advantage of PerpetualWonder is its ability to handle sequential, long-horizon action inputs within a single scene. We compare with WonderPlay [21] and demonstrate this capability in Figure 5 and also Figure 1. WonderPlay exhibits severe accumulated errors as the interactions progress. For instance, after the shovel is rotated in the air, it fails to maintain its shape integrity and becomes unrealistically deformed upon insertion into the castle. This is a fundamental limitation of its design: the generative refinements from the last time window do not propagate back to update the physical particles in simulators for the next round of interaction. As a result, at the initialization of each new round, the gaussian primitives are reset to their original positions rather than the optimized positions according to the refined video. The discrepancy causes severe fracturing artifacts and disrupts the temporal continuity across time windows. The decoupling, combined with the reliance on single-view refinement, results in visual artifacts and unstable dynamics as errors compound.

In contrast, our VPP representation provides a bidirectional bridge, allowing the refined state $\mathcal{S}_T$ from the end of one round to become the corrected initial state $\mathcal{S}_0$ for the

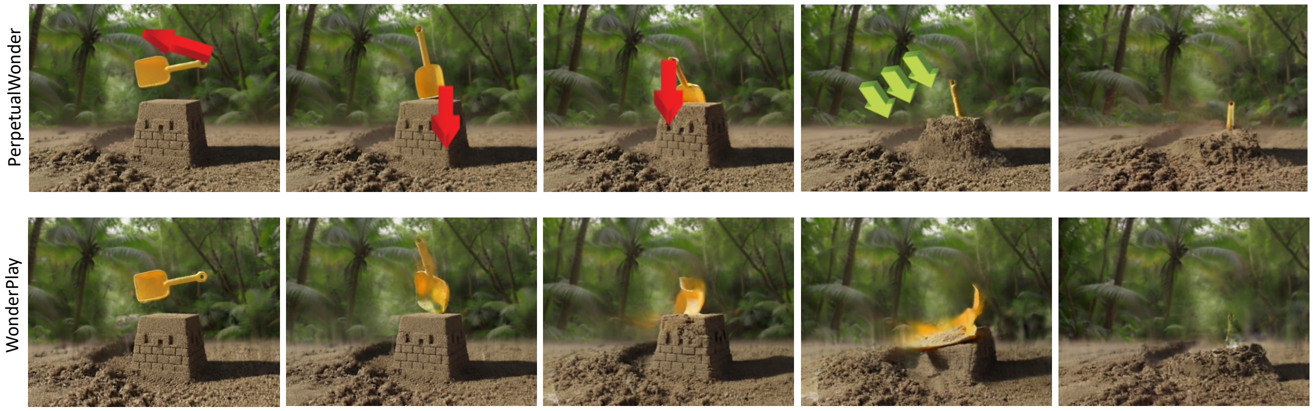

Figure 5. **Long-horizon actions** comparison between PerpetualWonder (top row) and WonderPlay (bottom row). For each method, the view changes across time, illustrating the four-round interaction results on a castle scene. The applied actions are overlaid on the top row.

Figure 6. **Ablation** on VPP representation. Top row is with our proposed VPP for the foreground. Bottom row shows using 3D gaussians from the standard Gaussian Splatting [18] optimization.

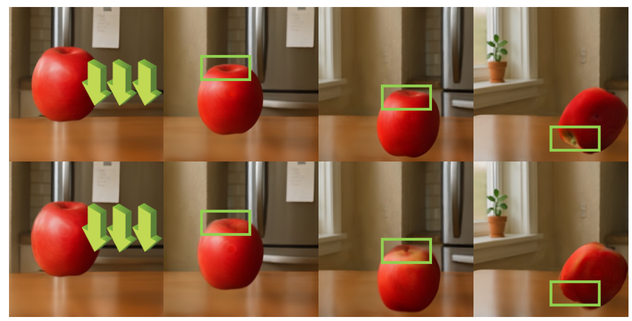

Figure 7. **Ablation** on progressive multi-view optimization. Top row shows the optimized scene using progressive optimization and the bottom row shows direct optimization results.

next round. As demonstrated in Figure 3, which provides more interactive 4D scenes, each with three rounds of interaction, PerpetualWonder can support sequential interactions for diverse object types such as elastic bodies, gases, and rigid objects, all while generating realistic physical motion.

4.2. Ablation Study

We perform an ablation study to assess the respective roles of our VPP representation and the progressive multi-view optimization strategy in generating plausible dynamics and maintaining multi-view consistency.

Inherent consistency and plausible dynamics from VPP. In the top row of Figure 6, we show the dynamic scene represented by VPP after multi-view optimization. The VPP representation constrains visual primitives $\{g_{j,k}\}$ to remain near their corresponding physics particle p_j. This binding mechanism accurately drives the visual primitives according to the dynamics from physical solvers. We contrast this with a variant (bottom row) that uses standard 3D gaussian primitives. When subjected to the same multi-view optimization, these unconstrained primitives just aim for minimizing the photometric loss, resulting in degenerate results with chaotic dynamics and visual artifacts.

Progressive optimization enhances the multi-view consistency. In Figure 7, we ablate our progressive optimization strategy (top row) against a direct optimization variant (bottom row). In direct optimization, video generators will not naturally generate perfectly consistent multi-view videos. For example, the generated video from the frontal view might hallucinate incorrect visual details (e.g., color or shape artifacts), while another view does not, leading to optimization conflicts. By using these inconsistent supervision signals to optimize the underlying scene at once, the representation becomes corrupted, which manifests as blurry textures and appearance flickering on the apple as time progresses. Our progressive approach successfully resolves this ambiguity, yielding a consistent 4D scene.

5. Conclusion

We introduce PerpetualWonder, a novel framework for long-horizon action-conditioned 4D scene generation from a single image. PerpetualWonder enables sequential interactions by unifying the physical and visual representations, and leveraging multi-view optimization for consistent updates. We demonstrate the superior performance of PerpetualWonder with diverse scenes and long-horizon actions.

Acknowledgments. This work is in part supported by NSF RI #2211258 and #2338203, ONR YIP N00014-24-1-2117, ONR MURI N00014-22-1-2740, Accenture, the Stanford Institute for Human-Centered AI (HAI), and the Magic Grant from the Brown Institute for Media Innovation.

References

[1] 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. 2

[2] 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. 2

[3] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*, 2025. 3

[4] Tim Brooks, Bill Peebles, et al. Video generation models as world simulators. OpenAI Technical Report, 2024. 1, 2, 3

[5] Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, et al. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13–23, 2025. 5, 6

[6] 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. 1, 3

[7] Tsai-Shien Chen, Chieh Hubert Lin, Hung-Yu Tseng, Tsung-Yi Lin, and Ming-Hsuan Yang. Motion-conditioned diffusion model for controllable video synthesis. *arXiv preprint arXiv:2304.14404*, 2023. 3

[8] Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. In *ICCV*, 2025. 6, S1

[9] Chen Geng, Yunzhi Zhang, Shangzhe Wu, and Jiajun Wu. Birth and death of a rose. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26102–26113, 2025. 2

[10] Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Yusuf Aytar, Michael Rubinstein, Chen Sun, et al. Motion prompting: Controlling video generation with motion trajectories. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1–12, 2025. 3

[11] Nate Gillman, Charles Herrmann, Michael Freeman, Daksh Aggarwal, Evan Luo, Deqing Sun, and Chen Sun. Force prompting: Video generation models can learn and generalize physics-based control signals. *arXiv preprint arXiv:2505.19386*, 2025. 3

[12] Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, et al. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–12, 2025. 6

[13] Haoyang Huang, Guoqing Ma, Nan Duan, Xing Chen, Changyi Wan, Ranchen Ming, Tianyu Wang, Bo Wang, Zhiying Lu, Aojie Li, et al. Step-video-ti2v technical report: A state-of-the-art text-driven image-to-video generation model. *arXiv preprint arXiv:2503.11251*, 2025. 3

[14] Tianyu Huang, Haoze Zhang, Yihan Zeng, Zhilu Zhang, Hui Li, Wangmeng Zuo, and Rynson WH Lau. Dreamphysics: Learning physics-based 3d dynamics with video diffusion priors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3733–3741, 2025. 1, 3

[15] Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson WH Lau, Wangmeng Zuo, and Chunchao Guo. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *arXiv preprint arXiv:2506.04225*, 2025. 1

[16] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. S1

[17] Chenfanfu Jiang, Craig Schroeder, Joseph Teran, Alexey Stomakhin, and Andrew Selle. The material point method for simulating continuum materials. *Acm siggraph 2016 courses*, 2016. 5

[18] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2, 4, 8

[19] Caterina Lacerra, Rocco Tripodi, Roberto Navigli, et al. Genesis: a generative approach to substitutes in context. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. ACL, 2021. 5, S1

[20] Long Le, Ryan Lucas, Chen Wang, Chuhao Chen, Dinesh Jayaraman, Eric Eaton, and Lingjie Liu. Pixie: Fast and generalizable supervised learning of 3d physics from pixels. *arXiv preprint arXiv:2508.17437*, 2025. 2

[21] 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. 2025. 2, 3, 4, 5, 6, 7, S1

[22] Huan Ling, Seung Wook Kim, Antonio Torralba, Sanja Fidler, and Karsten Kreis. Align your gaussians: Text-to-4d with dynamic 3d gaussians and composed diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8576–8588, 2024. 2

[23] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view syn-

thesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[24] Matthias Müller, Bruno Heidelberger, Matthias Teschner, and Markus Gross. Meshless deformations based on shape matching. *ACM transactions on graphics (TOG)*, 24(3):471–478, 2005. 5

[25] Matthias Müller, Bruno Heidelberger, Marcus Hennix, and John Ratcliff. Position based dynamics. *Journal of Visual Communication and Image Representation*, 18(2):109–118, 2007. 5

[26] Muyao Niu, Xiaodong Cun, Xintao Wang, Yong Zhang, Ying Shan, and Yinqiang Zheng. Mofa-video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model. In *European Conference on Computer Vision*, pages 111–128. Springer, 2024. 3

[27] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5865–5874, 2021. 2

[28] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021.

[29] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10318–10327, 2021. 2

[30] M Nomaan Qureshi, Sparsh Garg, Francisco Yandun, David Held, George Kantor, and Abhisesh Silwal. Splatsim: Zero-shot sim2real transfer of rgb manipulation policies using gaussian splatting. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6502–6509. IEEE, 2025. 1

[31] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 4, S1

[32] Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023. 2

[33] Jiawei Ren, Cheng Xie, Ashkan Mirzaei, Karsten Kreis, Ziwei Liu, Antonio Torralba, Sanja Fidler, Seung Wook Kim, Huan Ling, et al. L4gm: Large 4d gaussian reconstruction model. *Advances in Neural Information Processing Systems*, 37:56828–56858, 2024. 2

[34] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 3, 4, 5, 6, 7, S1

[35] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4

[36] Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, et al. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3

[37] 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. 2

[38] Xiyang Tan, Ying Jiang, Xuan Li, Zeshun Zong, Tianyi Xie, Yin Yang, and Chenfanfu Jiang. Physmotion: Physics-grounded dynamics from a single image. *arXiv preprint arXiv:2411.17189*, 2024. 2, 3

[39] HunyuanWorld Team, Zhenwei Wang, Yuhao Liu, Junta Wu, Zixiao Gu, Haoyuan Wang, Xuhui Zuo, Tianyu Huang, Wenhuan Li, Sheng Zhang, et al. Hunyuanworld 1.0: Generating immersive, explorable, and interactive 3d worlds from words or pixels. *arXiv preprint arXiv:2507.21809*, 2025. 1

[40] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 3, 6, 7

[41] Qinghe Wang, Yawen Luo, Xiaoyu Shi, Xu Jia, Huchuan Lu, Tianfan Xue, Xintao Wang, Pengfei Wan, Di Zhang, and Kun Gai. Cinemaster: A 3d-aware and controllable framework for cinematic text-to-video generation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–10, 2025. 3

[42] Yifan Wang, Peishan Yang, Zhen Xu, Jiaming Sun, Zhanhua Zhang, Yong Chen, Hujun Bao, Sida Peng, and Xiaowei Zhou. Freetimegs: Free gaussian primitives at anytime anywhere for dynamic scene reconstruction. In *CVPR*, 2025. 4

[43] Daniel Watson, Saurabh Saxena, Lala Li, Andrea Tagliasacchi, and David J Fleet. Controlling space and time with diffusion models. *arXiv preprint arXiv:2407.07860*, 2024. 2

[44] Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025. 2, 3, 6

[45] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20310–20320, 2024. 2

[46] 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. *arXiv:2411.18613*, 2024. 2

[47] Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. Draganything: Motion control for anything using entity representation. In *European Conference on Computer Vision*, pages 331–348. Springer, 2024. 3

[48] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4389–4398, 2024. 1, 3

[49] Zeyu Yang, Zijie Pan, Xiatian Zhu, Li Zhang, Jianfeng Feng, Yu-Gang Jiang, and Philip HS Torr. 4d gaussian splatting: Modeling dynamic scenes with native 4d primitives. *arXiv preprint*, 2024. 2

[50] 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. *arXiv preprint arXiv:2408.06072*, 2024. 3, 5

[51] Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. In *International Conference on Learning Representations (ICLR)*, 2024. 2

[52] Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes. In *ECCV*, 2024. 4, S1

[53] Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. Wonderjourney: Going from anywhere to everywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6658–6667, 2024. 5

[54] Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5916–5926, 2025. 5

[55] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In *CVPR*, 2017. 4, S1

[56] Yifei Zeng, Yanqin Jiang, Siyu Zhu, Yuanxun Lu, Youtian Lin, Hao Zhu, Weiming Hu, Xun Cao, and Yao Yao. Stag4d: Spatial-temporal anchored generative 4d gaussians. In *European Conference on Computer Vision*, pages 163–179. Springer, 2024. 2

[57] Bowen Zhang, Sicheng Xu, Chuxin Wang, Jiaolong Yang, Feng Zhao, Dong Chen, and Baining Guo. Gaussian variation field diffusion for high-fidelity video-to-4d synthesis. *arXiv preprint arXiv:2507.23785*, 2025. 2

[58] Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y. Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T. Freeman. PhysDreamer: Physics-based interaction with 3d objects via video generation. In *European Conference on Computer Vision*. Springer, 2024. 1, 3

[59] 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2063–2073, 2025. 3, 6

[60] Yuyang Zhao, Chung-Ching Lin, Kevin Lin, Zhiwen Yan, Linjie Li, Zhengyuan Yang, Jianfeng Wang, Gim Hee Lee, and Lijuan Wang. Genxd: Generating any 3d and 4d scenes. *arXiv preprint arXiv:2411.02319*, 2024. 2

[61] Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025. S1

[62] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024. *URL https://github. com/hpcaitech/Open-Sora*, 2024. 3

[63] Jensen Jinghao Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv preprint arXiv:2503.14489*, 2025. 3

PerpetualWonder: Long-Horizon Action-Conditioned 4D Scene Generation

Supplementary Material

A. Dense View Generation

Our goal is to generate dense, wide-range views of the underlying scene with the objects of interest from a single input image. The key here is to generate dense surrounding views with the scene content in the center and leverage the existing 3D reconstruction pipeline to reconstruct the underlying 3D scene. This approach is fundamentally different from the original image to 3D scene pipeline from WonderPlay [21]. The previous pipeline relies heavily on the single-view depth estimation and the alignment between the original image and the inpainted regions, resulting in a 3D scene representation that can only be viewed in a narrow baseline. On the contrary, our current approach leads to a full, complete 3D scene that supports rendering from arbitrary viewpoints.

For dense view generation, we employ GEN3C [34], which is capable of generating 3D-consistent multi-view videos from a single image. GEN3C is trained to generate multi-view videos of the underlying static scene, making it suitable for scene initialization. However, the vanilla GEN3C requires the generation to start from the input view. This presents a new challenge: generating a single, wide-angle (180° viewpoint change) camera trajectory often leads to inconsistent artifacts as the view deviates too far from the source (i.e., input image).

To acquire a dense set of novel views while maintaining consistency, we split the required camera viewpoint changes into two separate trajectories: "arc left" and "arc right". Both trajectories originate from the input view and rotate in different directions, with a 90° viewpoint change for each trajectory. We then generate a video for each trajectory and aggregate all the frames, forming a final, wide and consistent set of partial orbital views for the scene.

These generated dense views are further leveraged by our scene reconstruction pipeline for underlying 3D scene reconstruction, and also combined with SAM2 [31] and Gaussian Grouping [52] for foreground object segmentation.

B. Object Mesh Generation

Our pipeline further exploits TSDFusion [55] to reconstruct the mesh for the foreground objects. However, we empirically find that this approach is suitable for highly deformable materials like fluid and granular objects, but for rigid objects, the texture and geometry artifacts from incomplete TSDFusion reconstruction can pose severe challenges for the video generation model and lead to unexpected hallucination during the video refinement process.

To enhance the robustness of meshes for rigid objects, we exploit additional steps to improve the reconstruction quality for objects of this material. Specifically, we introduce Hunyuan3D [61], a powerful object-level image to 3D model. We directly leverage Hunyuan3D to generate the object mesh for rigid body objects, resulting in a complete surface mesh with much better quality compared to simple TSDFusion reconstruction, especially in the back regions.

Unlike the non-robust depth alignment and manual object placement steps in WonderPlay, we can further benefit from our aforementioned dense view generation pipeline and automatically position this generated mesh into the scene. Leveraging our synthesized multi-view images, we optimize the 6-DoF pose and scale by minimizing the projection error with respect to the surrounding views, ensuring it is perfectly aligned within the 3D scene.

C. Physical Simulation Parameters

The forward physics pass employs various solvers, and each solver requires specific physical parameter settings for reasonable simulation [19]. We provide a comprehensive list of these parameters, along with their default values, in Table S1. In practice, following the common practice in WonderPlay, these parameters are initially estimated using a Vision-Language Model [16] and are subject to optional manual fine-tuning to ensure physically plausible simulation results.

D. Evaluation metric details

To assess the scene quality, we render all generated scenes along a predefined camera trajectory and evaluate them using metrics from WorldScore [8]. Specifically, we use rule-based metrics to validate camera controllability and 3D consistency. We also include the imaging metric to assess general per-frame visual quality. To evaluate the plausible physical dynamics, we conducted a user study with 350 participants for this aspect. We employed a Two-alternative Forced Choice (2AFC) protocol, asking each participant to evaluate 10 scenes. For each scene, participants were given a multi-step interaction description and viewed a side-by-side, randomly ordered video comparison of our method and a baseline. Participants selected the video that performed better on one of two criteria: physics plausibility (the correctness of the predicted motion in response to the action) and motion fidelity (the quality and naturalness of the generated motion).

E. Visual-physical aligned Particle Configuration

While our VPP representation generally supports binding multiple gaussian primitives to a single physics particle, forming a set of size K, in practice, we configure K and the gaussian scale adaptively based on material properties to ensure optimal visual-physical alignment:

- **Solid and Surface Materials (Rigid body, Cloth):** We set $K = 1$. In this configuration, the gaussian scale is initialized to match the particle size δ. This strict one-to-one mapping ensures that the visual appearance is tightly constrained by the physics simulation, effectively preventing visual artifacts such as ghosting or detachment during large deformations.
- **Volumetric and Emitter Materials (Gas, Liquid, Sand, Snow, Elastic object):** To adequately represent the volumetric expansion and semi-transparent nature of these materials, we set $K = 20$ to allow a single physics particle to cover a larger visual volume. Correspondingly, the VPP's gaussian scale is initialized to be smaller than the particle size, defaulting to 0.5δ, to represent fine-grained volumetric details within the particle's influence radius.

F. Isotropic Visual Primitives

We demonstrate the differences between isotropic and anisotropic visual primitives in Figure S1. We find that isotropic primitives help remove blurry artifacts in novel views, as they do not tend to overfit the input image.

G. Ablation on Radius

We show an ablation on particle radius in Figure S2. The default radius is set to δ. Within a reasonable range (from 0.25δ to 4δ), our results are robust to different values. However, an overly small radius ($\leq 0.01\delta$) leads to insufficient representational capability, and an overly large radius ($\geq 100\delta$) leads to instability in optimization.

H. Limitation Discussion and Failure Case.

We provide the detailed runtime breakdown in Table S2. PerpetualWonder is currently not real-time due to the backward optimization overhead. Figure S3 shows a failure case involving a hockey stick moving into the frame from out-of-view. In the middle, as the stick enters the field of view, it appears incomplete (a hockey stick should ideally be longer than it appears). It remains a future work to complete object geometry that is not seen in the input image.

Parameter	Default Value
General simulation	
Step time	$1e^{-3}$
Sub-steps number	10
Sampled particle size	$1e^{-2}$
Gravity	$(0, 0, -9.8)$
Rigid body solver	
friction coefficient	0.1
MPM solver	
Grid density	64
Elastic material Young's modulus	$3e^{5}$
Elastic material Poisson's ratio	0.2
Liquid material Young's modulus	$1e^{7}$
Liquid material Poisson's ratio	0.2
Granular material Young's modulus	$1e^{6}$
Granular material Poisson's ratio	0.2
Granular material Friction angle	45
PBD solver	
Cloth material stretch compliance	$1e^{-7}$
Cloth material bending compliance	$1e^{-5}$
Smoke material viscosity coefficient	0.1

Table S1. Simulation parameters and default values

Stage	**Initialization**	**Forward Pass**	**Backward Opt.**	**Total (1st Loop)**
Time	~8 min	<1 min	~7 min	~16 min

Table S2. Runtime Analysis.

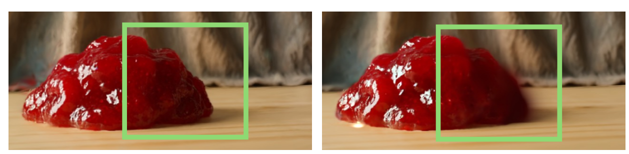

Figure S1. Comparison of isotropic and anisotropic primitives in novel view synthesis.

Figure S2. Ablation on radius.

Figure S3. Failure case in generating unseen geometry.