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.

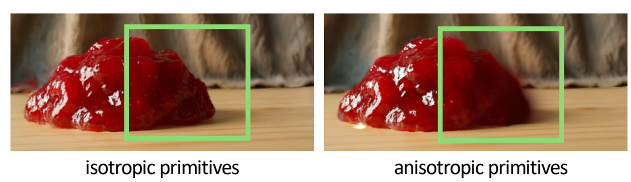

Figure S1. Comparison of isotropic and anisotropic primitives in novel view synthesis.

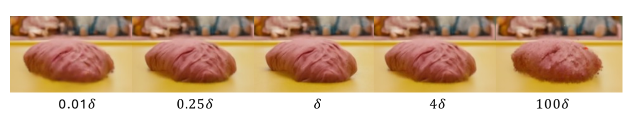

Figure S2. Ablation on radius.

Figure S3. Failure case in generating unseen geometry.