DreamGen: Real-Time Interactive Generation of 4D Dreamscapes

Anonymous ACL submission

Abstract

Recent advancements in visual generative models have substantially broadened the capabilities of scene synthesis across modalities such as video, 3D, and 4D environments which have significantly enhanced the application in various domains. Despite this progress, most existing systems treat scenes in isolation, lacking long-range spatial-temporal coherence and interactive control mechanisms. These shortcomings lead to the lack of interactivity and composability, limiting their potential in scenarios such as immersive entertainment and education. To address this, we introduce DreamGen, a novel unified framework designed to transform a single panoramic image into a fully interactive, panoramic 4D world. DreamGen operates through an integrated three-stage pipeline: First, it achieves view-consistent 3D reconstruction via Gaussian Splatting, employing monocular depth estimation and diffusion-based inpainting to enrich and complete the scene; next, it simulates continuous camera trajectories to ensure geometric and temporal consistency; finally, it combines these outputs within a real-time, event-driven Supersplat renderer to facilitate dynamic editing and immersive exploration. Extensive experiments on the comprehensive WorldScore benchmark demonstrate DreamGen's superior performance, outperforming existing state-of-the-art methods in controllability, visual fidelity, and motion dynamics. Our approach not only establishes new standards in interactive and coherent 4D world generation but also opens promising avenues for applications in immersive entertainment, embodied AI, and advanced simulation scenarios.

1 Introduction

Recent advances in visual generative models have significantly expanded the frontier of scene synthesis, enabling high-quality generation across modalities such as video, 3D, and 4D scenes (Wang et al., 2024; Xiao et al., 2025; Fremont et al., 2019; Li et al., 2024b; Yang et al., 2024a). These developments have laid the groundwork for the broader task of world generation—the construction of large-scale, coherent, and interactive environments composed of multiple, diverse scenes (Partarakis and Zabulis, 2024; Bruce et al., 2024; Park et al., 2023). World generation holds immense potential for applications in simulation, embodied AI, education, and immersive entertainment. However, current generative systems largely focus on isolated scene synthesis and fall short of producing temporally and spatially consistent multi-scene worlds with interactive capabilities. Bridging this gap requires not only new algorithmic frameworks, but also a deeper understanding of the unique challenges that world generation entails (Ha and Schmidhuber, 2018; Wu et al., 2023; Wang et al., 2025; Zhang et al., 2024).

A central challenge in world generation lies in synthesizing temporally coherent, spatially controllable, and visually diverse sequences that reflect a structured progression of scenes. Unlike traditional video or 3D scene generation, world generation requires models to reason over inter-scene dependencies, adhere to explicit spatial layouts, and preserve dynamic consistency across time (Meixner, 2017; Yu et al., 2024a) Existing approaches often lack fine-grained layout control and fail to support interactive manipulation or realistic temporal transitions. (Duan et al., 2025; Huang et al., 2024; Chen et al., 2024b; Hong et al., 2022). Furthermore, most prior benchmarks focus on single-scene fidelity under fixed modalities—such as text-to-video or single-view 3D reconstruction—and do not capture the sequential, compositional, and multimodal nature of world generation (Duan et al., 2025).

To address these limitations, we propose DreamGen, a unified framework for generating interactive panoramic 4D worlds from a single input image. The framework bridges the gap between static scene reconstruction and dynamic world modeling

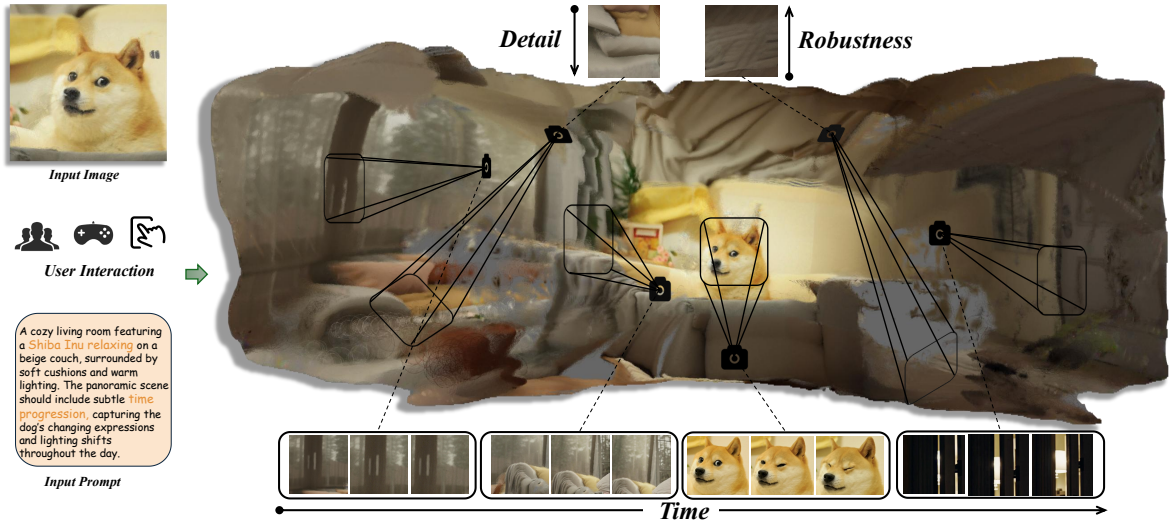

Figure 1: Illustrative input–prompt pair and evaluation axes. upper left: the single panoramic photograph fed to DreamGen. lower left: its accompanying natural-language prompt requesting. This pair serves as a running example for visualization results, where the generated 4D scene is assessed on the three WorldScore axes—*Controllability*, *Quality*, and *Dynamics*.

through a tightly coupled three-stage pipeline. First, the input image is lifted into a view-consistent 3D representation using Gaussian Splatting, guided by monocular depth estimation and diffusion-based inpainting to complete occluded or unseen regions. Building upon this static reconstruction, a continuous camera trajectory is simulated to generate temporally evolving views, where spatial-temporal reprojection ensures geometric and appearance consistency across frames. Finally, the resulting dynamic scene is rendered through an interactive, event-driven Supersplat renderer, enabling real-time playback and user-guided editing via timeline control. This holistic design facilitates coherent 4D world synthesis while supporting immersive and controllable scene exploration, as shown in figure 1.

We evaluate DreamGen on the WorldScore benchmark (Duan et al., 2025), which assesses 4D generation systems across dimensions of controllability, visual fidelity, and motion dynamics. DreamGen consistently outperforms competitive baselines, including LucidDreamer (Chung et al., 2023), CogVideoX (Yang et al., 2024b), and 4D-fy (Bahmani et al., 2024a), establishing new state-of-the-art results. These findings highlight DreamGen's capability to generate photorealistic, temporally coherent worlds while supporting fine-grained user interaction for immersive scene exploration. In summary, our contributions are three-fold:

- We propose DreamGen, a unified framework that synthesizes interactive panoramic 4D worlds from a single image, effectively bridging static reconstruction and dynamic scene modeling.
- We design a novel three-stage pipeline that combines monocular depth estimation and diffusion-based inpainting for 3D Gaussian Splatting, spatial-temporal reprojection for consistent sequence generation, and a real-time Supersplat renderer for user-controllable 4D visualization.
- We conduct extensive experiments on the WorldScore benchmark, where DreamGen achieves state-of-the-art performance across controllability, visual quality, and motion consistency, outperforming strong baselines such as LucidDreamer, CogVideoX, and 4D-fy.

2 Related Work

Single-View 3D Reconstruction. Single-view 3D reconstruction, positioned at the forefront of computer vision and graphics research, has been extensively explored and applied across various studies (Tatarchenko et al., 2019; Xue et al., 2022; Tono et al., 2024; Zheng et al., 2024). However, The task remains inherently ill-posed, as reconstructing a complete 3D structure from a single viewpoint necessitates reasoning about unseen regions (Liu et al., 2024b). Learning-based methods have become dominant due to their robustness and usability (Yang et al., 2023a). For example, Zhang et al. (Zhang et al., 2017) pioneered generalized single-view voxel reconstruction. Represent the

3D space as a grid of voxels, where each voxel indicates whether it's occupied or empty. Alternatively, 3D shapes can be represented as a set of points in space (Leberl et al., 2010; Chen et al., 2024a; Guo et al., 2020; Liang et al., 2024; Zhu et al., 2024). While both representations are simple and effective, they may lack detailed surface information. Pixel2Mesh(Yang et al., 2023a) generates 3D mesh models from single RGB images. This method represents 3D shapes as a collection of vertices, edges, and faces, providing a more structured and surface-aware representation. Moreover, implicit representations, such as Signed Distance Functions (Lin et al., 2020) and Neural Radiance Fields (Yu et al., 2023), facilitate the reconstruction of high-fidelity 3D geometries; the former excels in capturing intricate surface details, while the latter enables accurate novel view synthesis, including applications in facial avatar modeling. But implicit fields like SDFs or NeRFs give static, heavy models. DreamGen remedies these limits by combining monocular depth with diffusion inpainting to recover occluded regions, then converting the result into a lightweight 3D-Gaussian scene.

4D Scene Reconstruction. Moving beyond static reconstruction, 4D scene reconstruction focuses on capturing both spatial and temporal variations in dynamic scene (Li et al., 2024c; Weng et al., 2022; Yang et al., 2023a; Wu et al., 2024; Yu et al., 2023; Xu et al., 2024). Several recent concurrent studies (Li et al., 2024c; Luiten et al., 2024; Xie et al., 2024; Yang et al., 2024c, 2023b) have also demonstrated real-time rendering performance by integrating temporal coherence or time dependency into 3DGS. These methods either fail to produce significant and rapid action representations within the dataset (Li et al., 2024c) or are limited to generating only moderate-resolution outputs (Yang et al., 2023b). In contrast, 4K4DGen (Li et al., 2024a) demonstrates the ability to generate 360° panoramic omnidirectional dynamic scenes at 4K (4096 × 2048) resolution. DynamicScaler (Liu et al., 2024a) overcomes the limitations of 4K4DGen's range of motion for scalable panoramic dynamic scene synthesis with seamless motion capabilities.

Physics-based Interaction. While 4D scene reconstruction captures spatiotemporal variations, interactive modeling remains largely constrained to 3D due to the lack of explicit physical constraints in dynamic scenes (Weng et al., 2022; Sanchez-Gonzalez et al., 2020; Jiang et al., 2024). PhysGaussian (Xie et al., 2024), integrate physics-aware constraints into 3D Gaussian Splatting, enabling more structured and plausible scene dynamics. LucidDreamer (Chung et al., 2023) renders faster and in real-time with Gaussian Splatting, allowing users to interactively adjust the scene's perspective, lighting, and even local structure. Unlike the LucidDreamer, WonderWorld (Yu et al., 2024a) focuses on making 3D scene generation more interactive, accelerating scenario generation through FLAGS and reducing the time cost of extended scenarios. Existing methods either focus on 4D reconstruction without interactivity or constrain interaction to static 3D scenes. Our work integrates spatiotemporal modeling with physics-aware constraints, enabling real-time, interactive 4D scene manipulation while maintaining temporal coherence.

3 Preliminary

In this section, we describe the underlying principles of constructing a multi-view consistent 3D scene from a single panoramic image, following the LucidDreamer (Chung et al., 2023) pipeline. It reconstructs a navigable 3D scene from a *single* panoramic image by iteratively alternating *Dreaming* and *Alignment*. A monocular estimator first predicts a depth map D_0 for the input panorama I_0, and the image–depth pair is lifted into an initial point cloud

$$P_0 = \phi_{2\to3}\big([I_0, D_0], K, \Omega\big), \tag{1}$$

where K denotes camera intrinsics and Ω the pixel domain.

At iteration t, the current cloud P_t is rendered from a novel pose T_{t+1} to yield a colour–depth hint

$$\hat{I}_{t+1} = \phi_{3\to2}\big(P_t, K, T_{t+1}\big). \tag{2}$$

A diffusion network hallucinates the missing regions,

$$I_{t+1} = S(\hat{I}_{t+1}), \tag{3}$$

and a depth estimator provides

$$\hat{D}_{t+1} = D(I_{t+1}). \tag{4}$$

A global scale d_{t+1}, obtained via an ℓ_1 fit on overlapping rays, aligns the depth as

$$D_{t+1} = d_{t+1}\,\hat{D}_{t+1}. \tag{5}$$

The inpainted view is lifted into 3D

$$\hat{P}_{t+1} = \phi_{2\to3}\big([I_{t+1}, D_{t+1}], K, T_{t+1}\big), \tag{6}$$

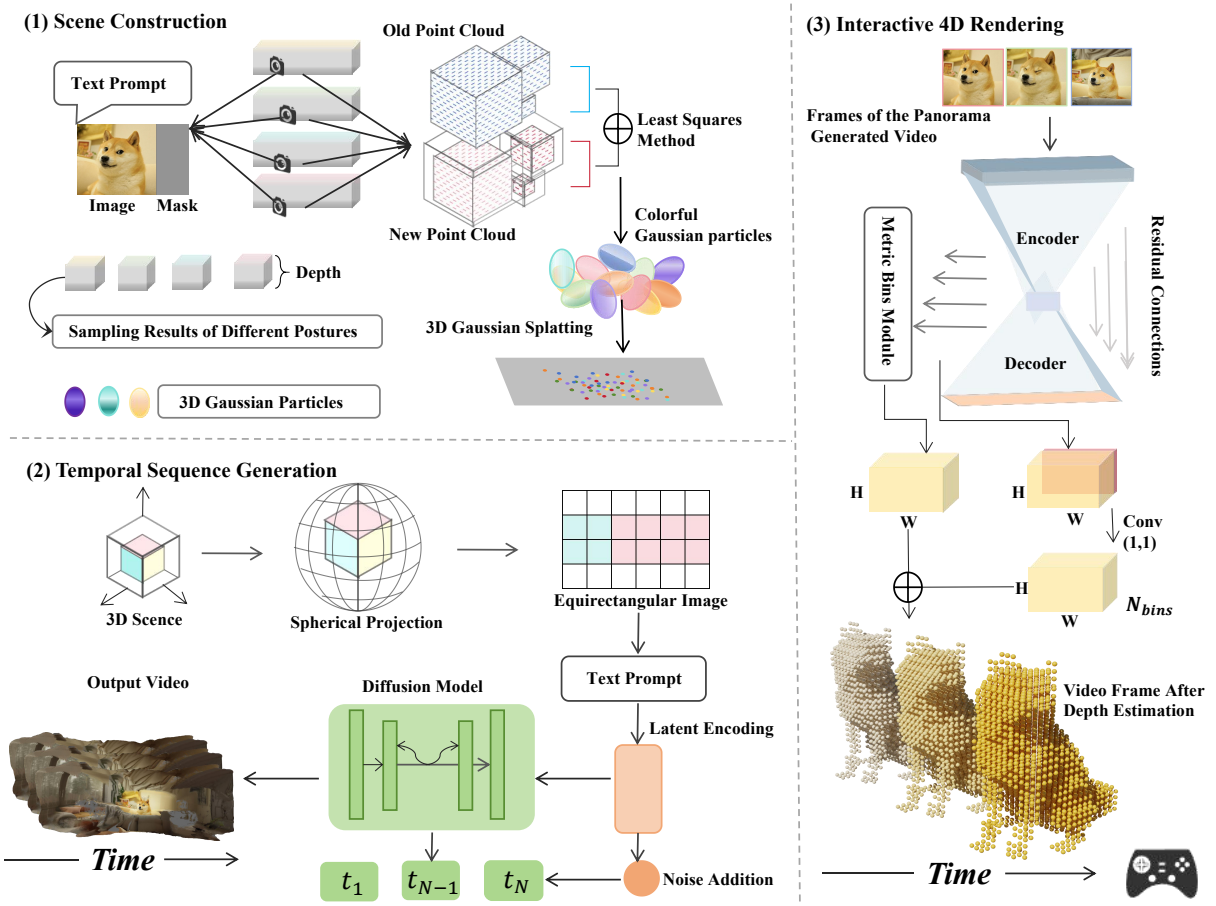

Figure 2: DreamGen pipeline overview. (1) Scene construction: a single panoramic image is lifted into a view-consistent 3D Gaussian scene via LucidDreamer, leveraging monocular depth estimation and diffusion-based inpainting for completeness. (2) Temporal sequence generation: an image-to-video model synthesizes a dynamic panorama, which is then decomposed into RGB–depth frames to form a temporally coherent sequence. (3) Interactive 4D rendering: the frame sequence is streamed to a real-time renderer, allowing users to explore and edit the evolving scene along the time axis.

then snapped to nearby rays of P_t to form $W(\hat{P}_{t+1})$. The scene is updated by

$$P_{t+1} = P_t \cup W(\hat{P}_{t+1}). \tag{7}$$

Repeating this Dreaming–Alignment cycle progressively densifies and stabilises the geometry, after which anisotropic Gaussians are optimised for photorealistic rendering.

4 Method

Taking a single panoramic image as input, the goal of DreamGen is to generate a panoramic 4D environment capable of interacting with humans. The pipeline of DreamGen is broadly divided into three stages: scene construction, temporal sequence generation, and interactive 4D rendering. In the first stage, a view-consistent 3D Gaussian splatting scene is constructed by leveraging monocular depth estimation and diffusion-based inpainting to expand the initial point cloud and enhance scene completeness. In the second stage, a temporal sequence is synthesized by projecting the 3D scene into a panoramic representation, from which individual frames are dynamically reconstructed into geometry-aware 3D structures to maintain spatial-temporal consistency. Finally, in the third stage, the reconstructed 4D scene undergoes Gaussian splat optimization and is integrated into an event-driven, immersive web-based rendering framework, enabling real-time exploration and manipulation, as shown in 2.

4.1 Scene Construction

Starting from P_0 (defined in Preliminary), we iterate Dreaming and Alignment as above to obtain a dense, view-consistent point cloud. After convergence we optimise 3-D Gaussians to obtain the

renderable scene used by subsequent stages.

4.2 Temporal Sequence Generation

Temporal sequence generation is systematically partitioned into two primary stages: View Rendering and Frame Reprojection and Playback Control. In the View Rendering stage, the 3D scene is projected onto the image plane along a continuously varying camera trajectory, thereby synthesizing a sequence of images that capture the scene from diverse perspectives. In the subsequent Frame Reprojection and Playback Control stage, each rendered frame is re-lifted into 3D space to recover its underlying geometric structure, while the temporal interval between successive frames serves as a critical parameter that governs the playback speed and ensures both temporal resolution and visual continuity in the resulting video.

View Rendering. Once the densified 3D scene is obtained as the point cloud P_N, we generate a temporal sequence by rendering the scene from a continuously varying set of camera poses. Let $\{P(t)\}_{t\in T}$ denote a smoothly parameterized camera trajectory over the temporal domain T. For each time instance $t \in T$, the projection of a 3D point $\mathbf{p} \in P_N$ onto the image plane is defined by a perspective projection function:

$$\mathbf{q} := K\, P(t)\, [\mathbf{p}^\top,\, 1]^\top,$$

$$\pi(\mathbf{p}, K, P(t)) = \big(\mathbf{q}_1/\mathbf{q}_3,\ \mathbf{q}_2/\mathbf{q}_3\big).$$

where K is the fixed camera intrinsic matrix and $P(t)$ encapsulates the camera extrinsics at time t.

The rendered image $I(t)$ is then synthesized by aggregating the contributions of all points in P_N. Formally, for each pixel coordinate (u, v) in the image, we define:

$$I(t)(u, v) = \sum_{\mathbf{p}\in P_N} w(\mathbf{p}, u, v; P(t))\, C(\mathbf{p}),$$

where $C(\mathbf{p})$ represents the color of point $\mathbf{p}$, and $w(\mathbf{p}, u, v; P(t))$ is a weighting function—often derived from a Gaussian kernel or splatting function—that quantifies the contribution of $\mathbf{p}$ to the pixel (u, v) based on the distance between $\pi(\mathbf{p}, K, P(t))$ and (u, v).

To form a panoramic video, the continuous trajectory $P(t)$ is discretely sampled at time instances

$$t_i = t_0 + i\,\Delta t, \quad i = 0, 1, 2, \ldots,$$

where Δt determines the temporal resolution (i.e., frame rate) of the video. The resulting sequence of frames $\{I(t_i)\}_{i=0}^{\infty}$ captures the scene as viewed from gradually changing perspectives, ensuring both spatial fidelity and temporal coherence across the panoramic video, as shown in 3.

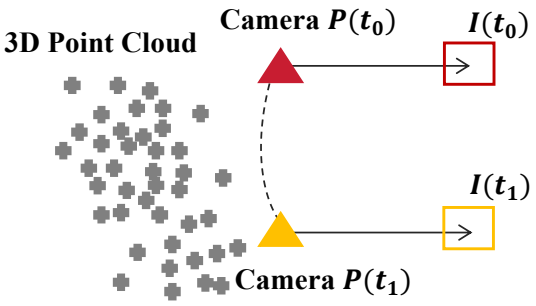

Figure 3: A smooth camera trajectory projects the reconstructed 3D point cloud P_N onto the image plane, rendering successive frames $I(t_0)$ and $I(t_1)$ from temporally varying viewpoints.

Frame Reprojection and Playback Control. Subsequent to view rendering, each generated frame $I(t)$ is reprojected back into 3D space to recover the underlying scene structure. Specifically, for every frame, a corresponding depth map $D(t)$ is estimated using a monocular depth estimation method. The lifting function $\phi_{2\to3}(\cdot)$ is then applied over the entire pixel domain Ω to reconstruct the set of 3D points:

$$P'(t) = \phi_{2\to3}\big([I(t),\, D(t)],\, K,\, \Omega\big).$$

This reprojection effectively retrieves the 3D geometry corresponding to each rendered view, ensuring that the synthesized imagery is consistent with the original spatial structure.

Furthermore, the temporal interval Δt between consecutive frames plays a crucial role in controlling the playback speed of the panoramic video. A smaller Δt results in a higher frame rate, leading to brisk transitions, whereas a larger Δt produces a slower, more gradual animation. By carefully selecting Δt, we can balance the need for smooth visual transitions with the fidelity of 3D reconstruction across the temporal sequence.

This two-stage process—comprising view rendering followed by frame reprojection—ensures that the dynamic panorama maintains both spatial accuracy and temporal coherence, thereby facilitating reliable interactive exploration of the synthesized 3D environment.

Table 1: Quantitative Evaluation of Generation Models

Models	Avg	Controllability				Quality			Dynamics		
		Camera Ctrl	Object Ctrl	Content Align	3D Consist	Photo Consist	Style Consist	Subjective Qual	Motion Acc	Motion Mag	Motion Smooth
3D Generation											
Allegro	51.97	24.84	57.47	51.48	70.50	69.89	65.60	47.41	54.39	40.28	37.81
Vchitect-2.0	38.47	26.55	49.54	65.75	41.53	42.30	25.69	44.58	33.59	33.81	21.31
SceneScape	35.51	84.99	47.44	28.64	76.54	62.88	21.85	32.75	-	-	-
Text2Room	43.47	94.01	38.93	50.79	88.71	88.36	37.23	36.69	-	-	-
LucidDreamer	49.28	88.93	41.18	75.00	90.37	90.20	48.10	58.99	-	-	-
WonderJourney	44.63	84.60	37.10	35.54	80.60	79.03	62.82	66.56	-	-	-
InvisibleStitch	42.78	93.20	36.51	29.53	88.51	89.19	32.37	58.50	-	-	-
WonderWorld	50.88	92.98	51.76	71.25	86.87	85.56	**70.57**	49.81	-	-	-
4D Generation											
4D-fy	32.10	69.92	55.09	0.85	35.47	1.59	32.04	0.89	22.22	22.88	**80.06**
DreamGen (Ours) (SVD)	63.00	72.57	42.70	58.90	70.64	94.94	55.89	46.97	**60.85**	**61.25**	65.29
DreamGen (Ours) (Hunyuan Video)	66.49	80.21	**68.15**	65.33	75.18	90.42	80.25	55.16	50.07	45.11	55.03
DreamGen (Ours) (Wan 2.1)	**68.14**	93.12	55.04	**76.15**	**91.01**	**94.95**	70.13	65.23	45.17	40.26	50.38

4.3 interactive 4D rendering

The interactive 4D rendering module is built upon a robust, event-driven framework that integrates real-time editing with smooth temporal animation. In our system, this is achieved by coupling a Supersplat-based renderer with a timeline module that governs frame progression and playback control.

Real-Time Editing. The real-time editing module constitutes the interactive core of our 4D rendering system, enabling dynamic manipulation of 3D scenes through an event-driven architecture that facilitates seamless communication among system components. At the foundation of this module is a centralized event bus, which serves as the primary mechanism for real-time interaction management. Within this framework (Contributors, 2025), multiple event-handling modules are systematically integrated, including register Camera PosesEvents, registe r Editor Events, register Selection Events, and register Transform Handler Events. These modules collectively enable real-time adjustments to the virtual camera, selection of scene objects using various tools (e.g., rectangle, brush, polygon, lasso, and sphere), and transformations such as translation, rotation, and scaling. By leveraging this event-driven infrastructure, modifications are instantaneously propagated through the rendering pipeline, ensuring a highly responsive and interactive editing experience.

Timeline control. The timeline control component orchestrates the temporal evolution of the 4D scene by ensuring smooth, cyclic animation and precise playback control. A dedicated timeline module continuously updates the current time and frame index based on the elapsed time and a predefined frame rate. Specifically, let Δt denote the elapsed time between update ticks and f_r the frame rate. At each update, the current time is updated as:

$$t \leftarrow (t + \Delta t \times f_r) \bmod F,$$

where F is the total number of frames. The discrete frame index is then determined by:

$$i = mod\Big(\lfloor t \rfloor,\, F\Big),$$

which ensures that the frame index wraps around cyclically for continuous looping. Furthermore, the timeline module provides interfaces to adjust the total frame count, frame rate, and keyframe settings, thereby offering fine-grained temporal control that synchronizes with user-driven real-time editing. These features maintain high temporal resolution and visual continuity throughout the interactive 4D rendering environment.

5 Experiment

In Section 4, we have elaborated on the design principles and detailed methodology of our proposed DreamGen framework, encompassing scene construction, temporal sequence generation, and interactive 4D rendering. To validate the effectiveness of our approach, we conduct comprehensive experiments and analyses in this section.

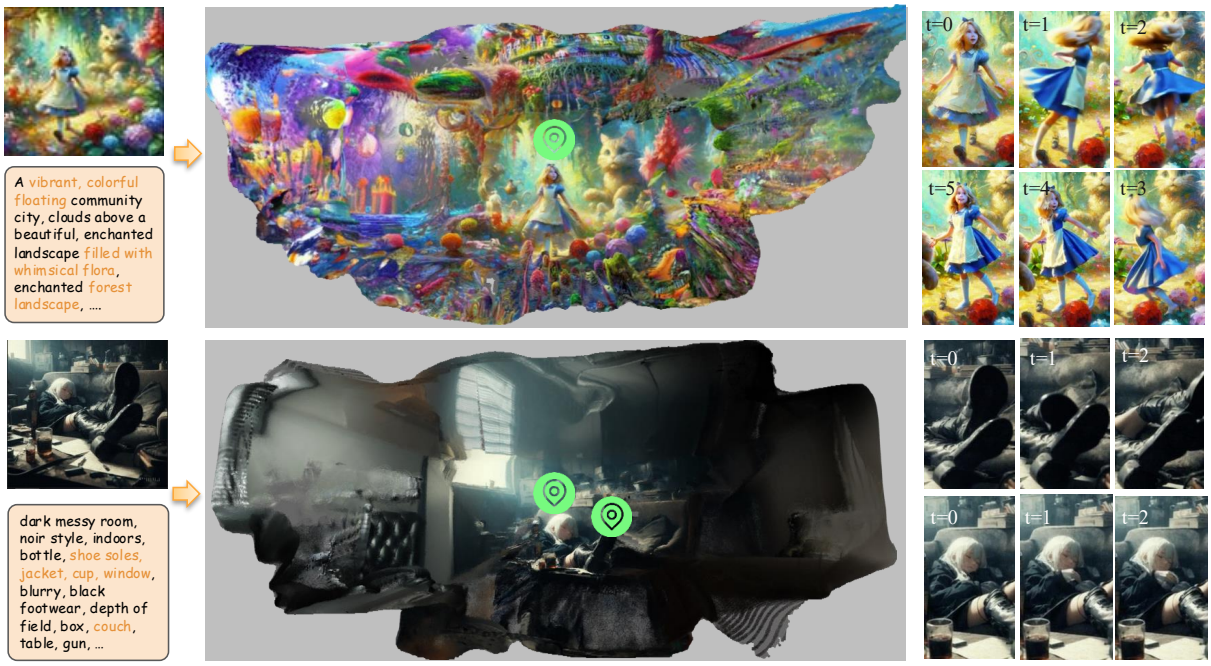

Figure 4: Visualization of DreamGen model outputs. Left panels show the input images and prompts. Center panels depict the generated 4D scenes, capturing detailed environments based on the inputs. Green markers indicate specific details that are dynamically animated, as shown in the sequences on the right panels.

5.1 Implementation Detail

Our method is implemented using the PyTorch framework, and all experiments are conducted on a server equipped with an NVIDIA TESLA A100 40G SXM4 GPU. Specifically, we utilize MiDaS (Ranftl et al., 2020) for monocular depth estimation, Stable Diffusion (Rombach et al., 2022) for diffusion-based image inpainting and scene expansion, and adopt the open-source implementation of LucidDreamer (Chung et al., 2023) for 3D Gaussian splatting. The interactive rendering module is built upon the SuperSplat (Contributors, 2025) framework, with the frontend interface developed using React and Three.js.

5.2 Quantitative Results

The results in Table 1 presents a comprehensive quantitative evaluation of our proposed DreamGen models compared to existing state-of-the-art (SOTA) 3D and 4D generation methods. Our DreamGen models consistently outperform previous methods across multiple evaluation metrics, demonstrating superior overall performance. Specifically, our best-performing variant, DreamGen (Wan 2.1), achieves the highest average score of 68.14, surpassing all other models by a significant margin. It notably excels in controllability metrics, achieving top scores in Camera Control (93.12), Content Alignment (76.15), and 3D Consistency (91.01). Additionally, DreamGen (Wan 2.1) demonstrates outstanding quality, obtaining the highest Photo Consistency score (94.95) and the second-highest Style Consistency (70.13) and Subjective Quality (65.23) scores. Our DreamGen (Hunyuan Video) variant also shows strong performance, particularly excelling in Object Control (68.15) and Style Consistency (80.25), indicating its capability to precisely control object placement and maintain stylistic coherence. In terms of dynamics, DreamGen (SVD) achieves the best performance, leading in Motion Accuracy (60.85) and Motion Magnitude (61.25), and ranking second in Motion Smoothness (65.29). This highlights its strength in generating accurate and smooth motion dynamics. Compared to existing 3D generation methods such as Allegro, LucidDreamer, and WonderWorld, our DreamGen models significantly improve upon both controllability and quality metrics. Furthermore, when compared to the existing 4D generation method (4D-fy), our models demonstrate substantial improvements across all metrics, particularly in Content Alignment, 3D Consistency, and Photo Consistency.

5.3 Visualization Results

Figure 4 illustrates the advanced capabilities of our DreamGen models in generating and animating complex scenes from textual descriptions. This fig-

ure showcases two distinct examples that highlight the versatility and effectiveness of our approach in handling diverse narrative contexts and visual styles.

In the first example, the input prompt describes *a vibrant, colorful floating community*. The central frame vividly depicts this imaginative setting, showing detailed structures suspended above an enchanted landscape. The sequence of animation frames on the right ($t = 0$ to $t = 5$) demonstrates the model's ability to generate consistent character motion while preserving the visual style and intricate details—an essential feature for animated storytelling and dynamic content creation.

In contrast, the second example is based on a noir-themed indoor scene described as *a dark, messy room*. The central image captures the somber atmosphere using dramatic lighting and shadowing techniques, accurately reflecting key elements such as the *bottle*, *shoe soles*, and *jacket*. The accompanying frame sequence ($t = 0$ to $t = 2$) highlights the model's capacity to simulate subtle scene dynamics, underscoring its strength in realistic 4D scene generation and nuanced visual storytelling.

Table 2: Ablation study highlighting the contributions of key components in DreamGen. Removing **ZoeDepth** significantly degrades all metrics, especially 3D consistency and subjective quality, demonstrating its critical role in depth-aware reconstruction. Omitting **Spherical Projection** leads to a notable drop in style and 3D consistency. Excluding **Video Rendering** severely harms 3D consistency despite maintaining photorealism.

Method	3D Consist.	Photo Consist.	Style Consist.	Subjective Qual.
DreamGen (Ours) (Wan 2.1)	76.15	91.01	94.95	70.13
w/o ZoeDepth	47.21	81.56	61.24	22.56
w/o Spherical Projection	66.21	90.66	85.26	70.01
w/o Video Rendering	27.53	90.81	92.10	70.34

5.4 Ablation Studies

The ablation results in Table 2, underscore the critical roles of specific components in our DreamGen. Removing ZoeDepth led to a significant drop in 3D consistency (from 76.15 to 47.21) and subjective quality (from 70.13 to 22.56), highlighting its importance in depth perception and overall aesthetic appeal. The absence of Spherical Projection slightly decreased photo consistency (from 91.01 to 90.66), suggesting its contribution to photorealistic rendering, albeit less critical than ZoeDepth. Similarly, omitting Video Rendering slightly affected subjective quality (from 70.13 to 70.34), indicating its role in enhancing dynamic visual content. These results demonstrate that each component is vital for maintaining the high quality and consistency of the generated images, confirming their collective contribution to the model's state-of-the-art performance. The removal of ZoeDepth introduces substantial degradation in both geometric integrity and texture consistency. As visualized in Figure 5, its absence leads to collapsed geometry, distorted surfaces, and fragmented regions, particularly in areas requiring fine depth reasoning such as lava contours and mountainous boundaries. Without ZoeDepth, the model fails to infer accurate depth from monocular cues, resulting in flattened structures and ambiguity.

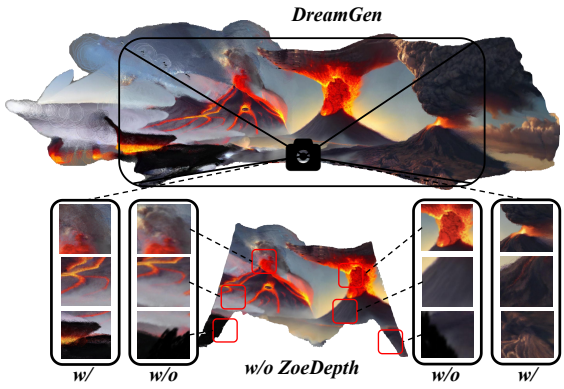

Figure 5: Visual comparison of 4D scene reconstruction quality with and without the integration of ZoeDepth in DreamGen. The panoramic volcanic scene reconstructed by DreamGen is shown at the top, with close-up regions highlighting structural fidelity and texture consistency. Red boxes mark significant artifacts or distortions when ZoeDepth is removed. The left and right columns display detailed patches for models w/ and w/o ZoeDepth, demonstrating improvements in geometry and texture realism when depth estimation is applied.

6 Conclusion

We present DreamGen, a unified framework for generating interactive panoramic 4D worlds from a single image. Through a three-stage pipeline, including scene construction, temporal sequence generation, and interactive rendering, DreamGen achieves state-of-the-art performance across controllability, quality, and dynamics on the WorldScore benchmark. Extensive experiments and visualizations demonstrate its ability to generate coherent, editable, and immersive 4D scenes, opening new directions for world modeling and embodied AI applications.

Limitation

Computational Overhead. Although DreamGen delivers state-of-the-art controllability and scene quality, its pipeline is still highly time-consuming. The iterative monocular–to–multi-view depth refinement and diffusion-based inpainting require multiple forward passes for every novel viewpoint, and the subsequent Gaussian-splat optimisation further extends the overall runtime. Such heavy computation limits the method's applicability in latency-sensitive scenarios such as AR/VR streaming or on-device content creation.

Real-Time Responsiveness. DreamGen's Supersplat renderer sustains interactive frame rates for moderately complex worlds, but performance degrades sharply as the point budget or shader complexity increases. In densely occluded outdoor scenes we observe noticeable input–display lag that impedes fine-grained camera control and object manipulation. While aggressive LOD pruning and foveated rendering can mitigate slowdowns, they risk introducing popping artefacts and degrading peripheral fidelity.

Perceptual Artefacts. Despite the LucidDreamer-based depth initialisation, failure modes persist in regions with transparent, specular, or textureless surfaces. These areas yield noisy depth estimates that propagate to the multi-view optimisation stage, manifesting as floating fragments or stretched splats in the final 4D reconstruction. Post-hoc bilateral filtering reduces noise but cannot fully recover fine geometry or eliminate dis-occlusion ghosts.

Directions for Future Work. Efficient *single-pass depth hallucination*, tensor-core-aware splat optimisation, and adaptive streaming strategies could substantially lower the computational burden and boost runtime FPS. Incorporating *uncertainty-aware depth networks* and *geometry-consistent diffusion priors* offers a promising path to mitigate artefacts in challenging photometric conditions. Finally, expanding evaluation to open-world, multi-agent settings would provide a more comprehensive picture of DreamGen's strengths and limitations.

References

Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B. Lindell. 2024a. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B Lindell. 2024b. 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.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, and 1 others. 2024. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*.

Guangyan Chen, Meiling Wang, Yi Yang, Kai Yu, Li Yuan, and Yufeng Yue. 2024a. Pointgpt: Auto-regressively generative pre-training from point clouds. *Advances in Neural Information Processing Systems*, 36.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. 2024b. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. *Preprint*, arXiv:2401.09047.

Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. 2023. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv preprint arXiv:2311.13384*.

SuperSplat Contributors. 2025. Supersplat - 3d gaussian splat editor. Version 2.1.0, accessed March 7, 2025.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. 2025. Worldscore: A unified evaluation benchmark for world generation. *arXiv preprint arXiv:2504.00983*.

Paul Engstler, Andrea Vedaldi, Iro Laina, and Christian Rupprecht. 2024. Invisible stitch: Generating smooth 3d scenes with depth inpainting. *arXiv preprint arXiv:2404.19758*.

Daniel J Fremont, Tommaso Dreossi, Shromona Ghosh, Xiangyu Yue, Alberto L Sangiovanni-Vincentelli, and Sanjit A Seshia. 2019. Scenic: a language for scenario specification and scene generation. In *Proceedings of the 40th ACM SIGPLAN conference on programming language design and implementation*, pages 63–78.

Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. 2020. Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(12):4338–4364.

David Ha and Jürgen Schmidhuber. 2018. World models. *arXiv preprint arXiv:1803.10122*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. 2023. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7909–7920.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. 2022. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, and 1 others. 2024. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818.

Ying Jiang, Chang Yu, Tianyi Xie, Xuan Li, Yutao Feng, Huamin Wang, Minchen Li, Henry Lau, Feng Gao, Yin Yang, and 1 others. 2024. Vr-gs: A physical dynamics-aware interactive gaussian splatting system in virtual reality. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–1.

Franz Leberl, Arnold Irschara, Thomas Pock, Philipp Meixner, Michael Gruber, Set Scholz, and Alexander Wiechert. 2010. Point clouds. *Photogrammetric Engineering & Remote Sensing*, 76(10):1123–1134.

Renjie Li, Panwang Pan, Bangbang Yang, Dejia Xu, Shijie Zhou, Xuanyang Zhang, Zeming Li, Achuta Kadambi, Zhangyang Wang, Zhengzhong Tu, and 1 others. 2024a. 4k4dgen: Panoramic 4d generation at 4k resolution. *arXiv preprint arXiv:2406.13527*.

Xiaoyu Li, Qi Zhang, Di Kang, Weihao Cheng, Yiming Gao, Jingbo Zhang, Zhihao Liang, Jing Liao, Yan-Pei Cao, and Ying Shan. 2024b. Advances in 3d generation: A survey. *arXiv preprint arXiv:2401.17807*.

Zhan Li, Zhang Chen, Zhong Li, and Yi Xu. 2024c. Spacetime gaussian feature splatting for real-time dynamic view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8508–8520.

Dingkang Liang, Xin Zhou, Wei Xu, Xingkui Zhu, Zhikang Zou, Xiaoqing Ye, Xiao Tan, and Xiang Bai. 2024. Pointmamba: A simple state space model for point cloud analysis. *arXiv preprint arXiv:2402.10739*.

Chen-Hsuan Lin, Chaoyang Wang, and Simon Lucey. 2020. Sdf-srn: Learning signed distance 3d object reconstruction from static images. *Advances in Neural Information Processing Systems*, 33:11453–11464.

Jinxiu Liu, Shaoheng Lin, Yinxiao Li, and Ming-Hsuan Yang. 2024a. Dynamicscaler: Seamless and scalable video generation for panoramic scenes. *arXiv preprint arXiv:2412.11100*.

Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. 2024b. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36.

Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. 2024. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *2024 International Conference on 3D Vision (3DV)*, pages 800–809. IEEE.

Britta Meixner. 2017. Hypervideos and interactive multimedia presentations. *ACM computing surveys (CSUR)*, 50(1):1–34.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Nikolaos Partarakis and Xenophon Zabulis. 2024. A review of immersive technologies, knowledge representation, and ai for human-centered digital experiences. *Electronics*, 13(2):269.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *Preprint*, arXiv:1907.01341.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. 2020. Learning to simulate complex physics with graph networks. In *International conference on machine learning*, pages 8459–8468. PMLR.

Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. 2019. What do single-view 3d reconstruction networks learn? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3405–3414.

Alberto Tono, Heyaojing Huang, Ashwin Agrawal, and Martin Fischer. 2024. Vitruvio: conditional variational autoencoder to generate building meshes via single perspective sketches. *Automation in Construction*, 166:105498.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. 2023. Exploring clip for assessing the look and feel of images. In *AAAI*.

Lening Wang, Wenzhao Zheng, Yilong Ren, Han Jiang, Zhiyong Cui, Haiyang Yu, and Jiwen Lu. 2024. Occsora: 4d occupancy generation models as world simulators for autonomous driving. *arXiv preprint arXiv:2405.20337*.

Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, and 1 others. 2025. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, 133(5):3059–3078.

Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. 2022. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, pages 16210–16220.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 2024. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20310–20320.

Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. 2023. Daydreamer: World models for physical robot learning. In *Conference on robot learning*, pages 2226–2240. PMLR.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. 2025. Worldmem: Long-term consistent world simulation with memory. *Preprint*, arXiv:2504.12369.

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

Zhen Xu, Sida Peng, Haotong Lin, Guangzhao He, Jiaming Sun, Yujun Shen, Hujun Bao, and Xiaowei Zhou. 2024. 4k4d: Real-time 4d view synthesis at 4k resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20029–20040.

Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. 2022. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5036–5045.

Xianghui Yang, Guosheng Lin, and Luping Zhou. 2023a. Single-view 3d mesh reconstruction for seen and unseen categories. *IEEE transactions on image processing*, 32:3746–3758.

Xiuyu Yang, Yunze Man, Junkun Chen, and Yu-Xiong Wang. 2024a. Scenecraft: Layout-guided 3d scene generation. *Advances in Neural Information Processing Systems*, 37:82060–82084.

Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. 2023b. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. *arXiv preprint arXiv:2310.10642*.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, and 1 others. 2024b. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*.

Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. 2024c. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20331–20341.

Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T. Freeman, and Jiajun Wu. 2024a. Wonderworld: Interactive 3d scene generation from a single image. *arXiv:2406.09394*.

Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, and 1 others. 2024b. Wonderjourney: Going from anywhere to everywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6658–6667.

Wangbo Yu, Yanbo Fan, Yong Zhang, Xuan Wang, Fei Yin, Yunpeng Bai, Yan-Pei Cao, Ying Shan, Yang Wu, Zhongqian Sun, and 1 others. 2023. Nofa: Nerf-based one-shot facial avatar reconstruction. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–12.

Yinda Zhang, Mingru Bai, Pushmeet Kohli, Shahram Izadi, and Jianxiong Xiao. 2017. Deepcontext: Context-encoding neural pathways for 3d holistic scene understanding. In *Proceedings of the IEEE international conference on computer vision*, pages 1192–1201.

Yuanhan Zhang, Kaichen Zhang, Bo Li, Fanyi Pu, Christopher Arif Setiadharma, Jingkang Yang, and Ziwei Liu. 2024. Worldqa: Multimodal world knowledge in videos through long-chain reasoning. *arXiv preprint arXiv:2405.03272*.

Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. 2024. Occworld: Learning a 3d occupancy world model for autonomous driving. In *European conference on computer vision*, pages 55–72. Springer.

Qinfeng Zhu, Lei Fan, and Ningxin Weng. 2024. Advancements in point cloud data augmentation for deep learning: A survey. *Pattern Recognition*, page 110532.

A Appendix

A.1 Baseline

Text2Room Text2Room (Höllein et al., 2023), noted around March 2023, is an innovative method designed to generate room-scale, textured 3D meshes directly from a given text prompt, functioning effectively as an Image-to-Video (or rather, image-sequence-to-3D) system for 3D scene construction at resolutions like 512x512. It achieves this by utilizing pre-trained 2D text-to-image models to synthesize a sequence of images from various viewpoints, which are then cohesively lifted into a consistent 3D scene representation through a combination of monocular depth estimation and a text-conditioned inpainting model.

LucidDreamer LucidDreamer (Chung et al., 2023) is a sophisticated framework, with research published around November 2023, for the domain-free generation of high-fidelity 3D Gaussian Splatting scenes from either a single text prompt or an image, often operating with 512x512 image inputs/outputs for its 2D components. It uniquely employs a recursive "Dreaming and Alignment" methodology, leveraging large-scale diffusion models for creating multi-view consistent images that are subsequently elevated to 3D. A key innovation is its use of Interval Score Matching (ISM) to produce detailed and realistic 3D models, effectively mitigating the over-smoothing issues prevalent in earlier Score Distillation Sampling (SDS) based methods.

WonderJourney WonderJourney (Yu et al., 2024b) is an Image-to-3D Scenes model, highlighted in research around December 2023, and was notably evaluated as part of the comprehensive WorldScore benchmark for world generation. The model is particularly recognized for its strong emphasis on creating extensive and comprehensive virtual worlds from image inputs, positioning it as a significant contributor to advancements in large-scale, dynamic scene generation.

InvisibleStitch InvisibleStitch (Engstler et al., 2024) is an Image-to-3D Scenes model, with developments noted around April 2024, specifically mentioned for its application in the realm of 3D content generation, working with image inputs at resolutions like 512x512. While detailed public documentation on a uniquely named "InvisibleStitch" model can be limited, its context within 3D I2V evaluations suggests a focus on seamlessly creating or integrating elements within 3D scenes or videos derived from initial static images.

WonderWorld WonderWorld (Yu et al., 2024a), with research contributions noted around June 2024, is an Image-to-Video 3D scene generation model that was also a subject of evaluation in the WorldScore benchmark, handling inputs/outputs around 512x512. It is particularly distinguished by its interactive capabilities and its specialized focus on generating immersive 3D environments originating from single input images, thereby playing a key role in assessing diverse aspects of world generation technologies.

4D-fy 4D-fy (Bahmani et al., 2024b) is a cutting-edge Text-to-4D generation technique that synthesizes dynamic 3D scenes (4D content) from textual prompts. It introduces an innovative "Hybrid Score Distillation Sampling" (Hybrid SDS) method, which strategically blends supervisory signals from multiple pre-trained diffusion models—including Text-to-Image (T2I), 3D-aware T2I, and Text-to-Video (T2V) models—through an alternating optimization process. This approach is designed to achieve state-of-the-art 4D scenes characterized by compelling visual appearance, robust 3D structure, and naturalistic motion.

A.2 Experimental Details

All experiments were executed on a single NVIDIA TESLA A100 40GB GPU (Ubuntu 22.04, PyTorch 2.0.1 compiled for CUDA 11.6). We fixed the random seed to 42 for `torch`, `numpy`, and Python's `random` module, and enabled deterministic cuDNN kernels to ensure bit-wise reproducibility. Throughout the pipeline we used a DDIM sampler with 50 diffusion steps. Scene construction employs a "*look-around*" camera sweep, while temporal synthesis follows a "*back_and_forth*" dolly; both trajectories are sampled at 30 fps.

A.3 WorldScore Evaluation Metrics

We adopt the WorldScore benchmark (Duan et al., 2025) to evaluate world generation performance across 3D, 4D, image-to-video (I2V), and text-to-video (T2V) paradigms. The benchmark assesses models along three main dimensions: controllability, quality, and dynamics, via ten specific metrics.

A.3.1 Controllability Metrics

Camera Controllability: Measures the deviation of the generated camera trajectory from the reference:

$$e_{\text{camera}} = \sqrt{e_\theta \cdot e_t} \tag{8}$$

where e_θ and e_t are the scale-invariant rotational and translational errors.

Object Controllability: Computes the success rate of open-set object detection using entities extracted from the next-scene prompt N. The model is rewarded if the mentioned objects appear in the generated scene.

Content Alignment: Evaluated by computing the CLIP score (Hessel et al., 2021) between the full prompt N and the generated video frames.

A.3.2 Quality Metrics

3D Consistency: Uses DROID-SLAM to calculate reprojection error between co-visible pixels across frames:

$$e_{\text{reproj}} = \frac{1}{|V|} \sum_{(i,j)\in V} \left\| p_{ij}^{*} - \Pi(P_{ij}) \right\|_2^2 \quad (9)$$

where p_{ij}^{*} is the observed 2D projection, and $\Pi(\cdot)$ is the camera projection function.

Photometric Consistency: Detects texture flickering using optical flow-based Average End-Point Error (AEPE):

$$e_{\text{photo}} = \frac{1}{N} \sum_{i=1}^{N} \left\| p_{A,i} - p'_{A,i} \right\|_2^2 \quad (10)$$

where $p'_{A,i}$ is the position tracked via backward optical flow from frame B.

Style Consistency: Computes visual style difference between the first and last frame using Gram matrix difference:

$$e_{\text{style}} = \|G_1 - G_T\|_F \quad (11)$$

Subjective Quality: An ensemble metric combining CLIP-IQA+ and CLIP-Aesthetic scores (Wang et al., 2023), selected to best align with human preference via a 200-participant study.

A.3.3 Dynamics Metrics

Motion Accuracy: Evaluates whether motion occurs in the intended dynamic region:

$$s_{\text{motion-acc}} = \max(F \odot M) - \max(F \odot \overline{M}) \quad (12)$$

where F is the optical flow magnitude and M is the mask for the target dynamic region.

Motion Magnitude: Captures the strength of motion by computing median flow magnitude across frames:

$$s_{\text{motion-mag}} = \text{median}(F) \quad (13)$$

Motion Smoothness: Assessed by interpolating dropped frames and comparing to ground truth using a combination of MSE, SSIM, and LPIPS scores.

A.3.4 Score Normalization

Each raw metric score s is normalized to the range $[0, 1]$ using:

$$s_{\text{norm}} = \left\langle \alpha \cdot \frac{s - b_{\min}}{b_{\max} - b_{\min}} + (1 - \alpha) \right\rangle \quad (14)$$

where $\alpha = 1$ for metrics where higher is better, and $\alpha = -1$ otherwise. $\langle \cdot \rangle$ denotes clipping to $[0, 1]$.

A.4 More Results

In this section, we present additional results showcasing the capabilities of our proposed method in generating 4D scenes from a single-view image prompt. As illustrated in Figure 6, the top sequence captures the temporal dynamics at a fixed spatial viewpoint, while the bottom sequence displays spatial variations across different viewpoints. These frames are jointly fused to reconstruct a temporally evolving 3D scene, forming a unified 4D representation. The central visualization demonstrates the resulting spatio-temporal scene, enabling immersive exploration along both time and space axes.

As illustrate in Figure 7, We two distinct applications of our 4D scene generation method, demonstrating its versatility and effectiveness across different environments and themes. In the top example, our method has generated a 4D interactive holiday living room scene from a single input image. The panoramic image was transformed to include a cozy scene that features a decorated Christmas tree, chairs, and ambient lighting. The user interaction pathway allows for navigation through the scene over time, showcasing how the environment evolves and reacts to simulated changes in viewpoint and lighting, enhancing the immersive experience. The bottom example showcases a dramatically different scenario—a highly detailed 4D rendered scene of a volcanic eruption. Starting from an input image of a mountain, our method dynamically models lava flowing down the slopes, with realistic smoke, ash, and backlighting effects that reflect the sunset illumination. This example highlights the method's capability to handle complex natural phenomena and render them with cinematic quality. The temporal dimension is particularly emphasized here, with the sequence of frames showing the progression of the eruption, offering

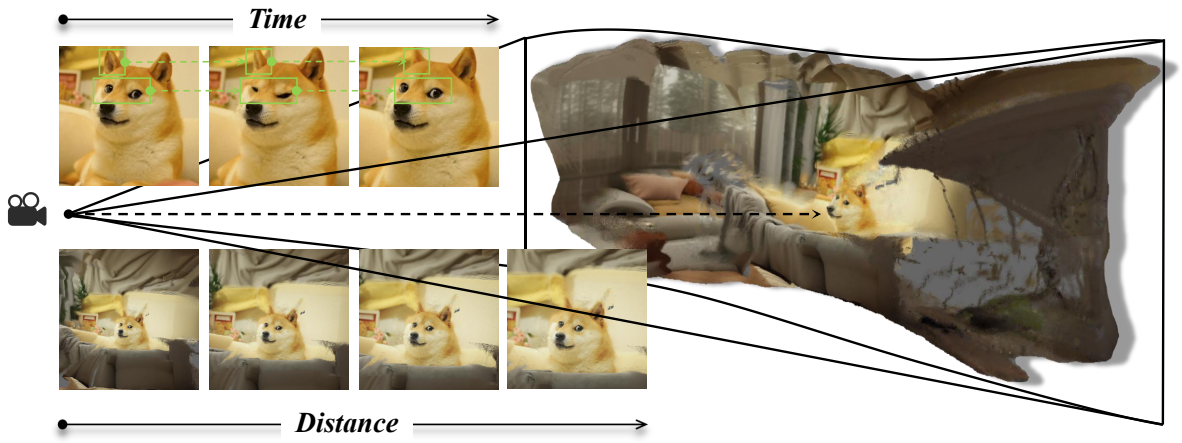

Figure 6: Generated 4D scene from a single-view image prompt. The top sequence illustrates temporal dynamics captured at a fixed spatial viewpoint, while the bottom sequence shows spatial variations across different viewpoints. These frames are jointly fused to reconstruct a temporally evolving 3D scene, forming a unified 4D representation. The center visualization demonstrates the resulting spatio-temporal scene, enabling immersive exploration along both time and space axes.

a compelling visualization of dynamic geological events.

B Reproducibility Statement

We used AI assistants (e.g., ChatGPT) for grammar correction and language refinement only. No content generation or experimental decision-making was done by AI tools.

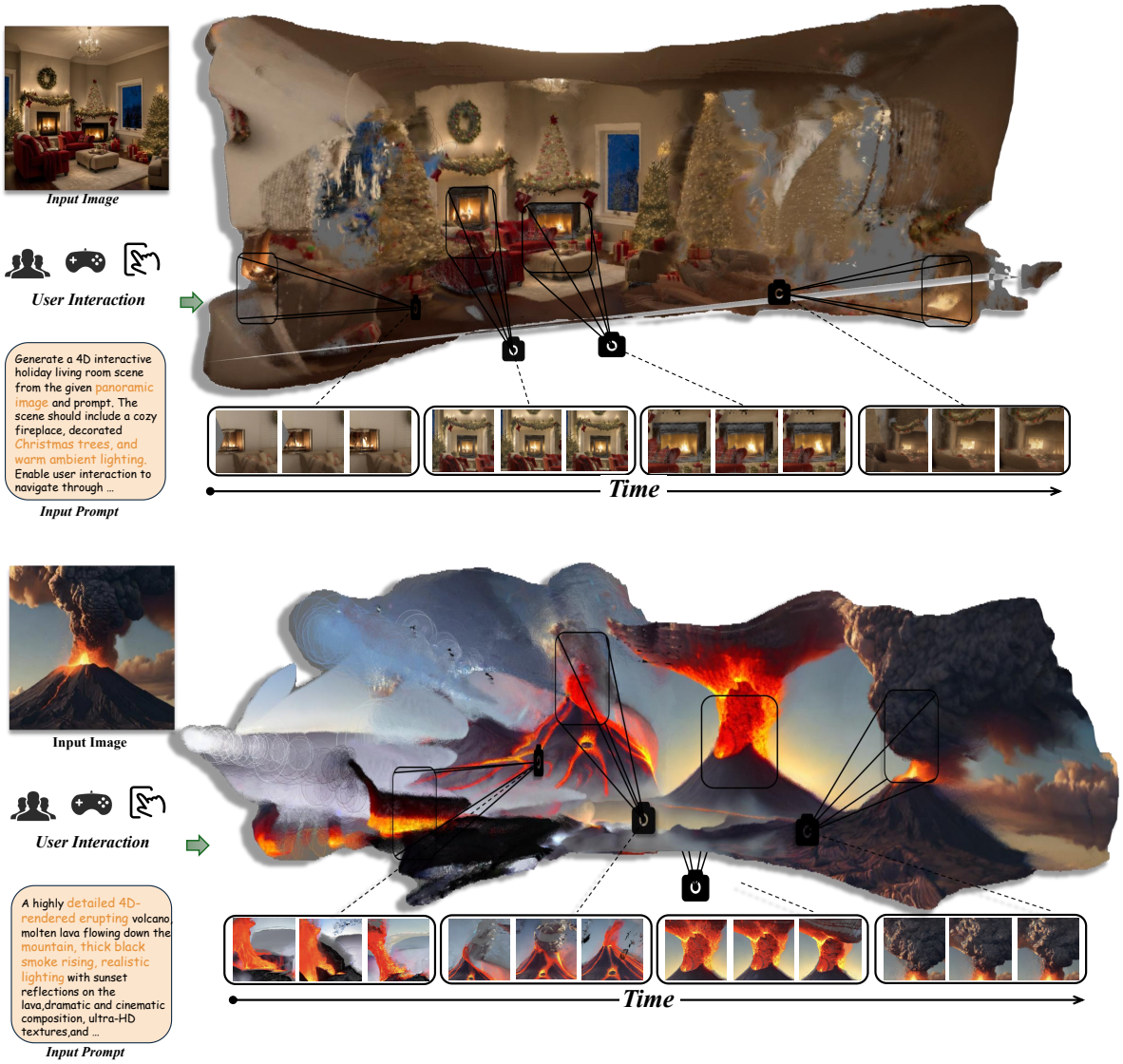

Figure 7: Demonstrating the versatility of our 4D scene generation method with two contrasting scenarios. Top: An interactive 4D holiday living room scene generated from a panoramic image, featuring a cozy fireplace, decorated Christmas trees, and dynamic ambient lighting, with user interaction paths allowing navigation through time. Bottom: A dynamic 4D rendering of a volcanic eruption, originating from a single image of a mountain, showcasing molten lava flows, realistic smoke and ash effects, and dramatic backlighting simulating sunset, with a timeline illustrating the progression of the eruption.