4K4DGen: Panoramic 4D Generation at 4K Resolution

Renjie Li[*1,4], Panwang Pan[*†1], Bangbang Yang[*1], Dejia Xu[*2], Shijie Zhou[3], Xuanyang Zhang[1], Zeming Li[1], Achuta Kadambi[3], Zhangyang Wang[2], Zhengzhong Tu[4], Zhiwen Fan[2]

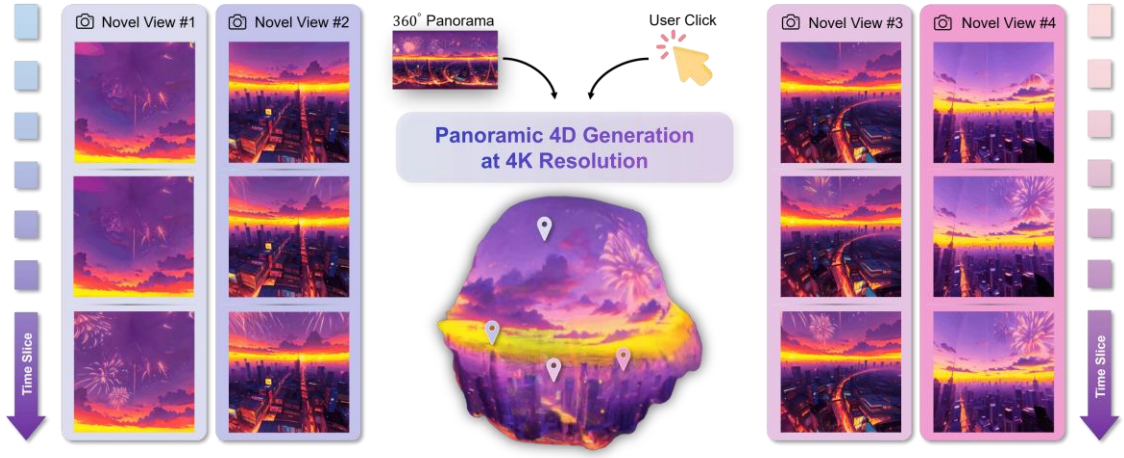
[1]Bytedance, [2]UT Austin, [3]UCLA, [4]TAMU, [*]Equal Contribution, [†]Corresponding Author

MOTIVATION

(i) The blooming of VR/AR demands for high-quality dynamic content creation for an immersive user experience.

(ii) Creating a consistent panoramic video at 4K resolution from a static 360° panorama and a user-defined target region.

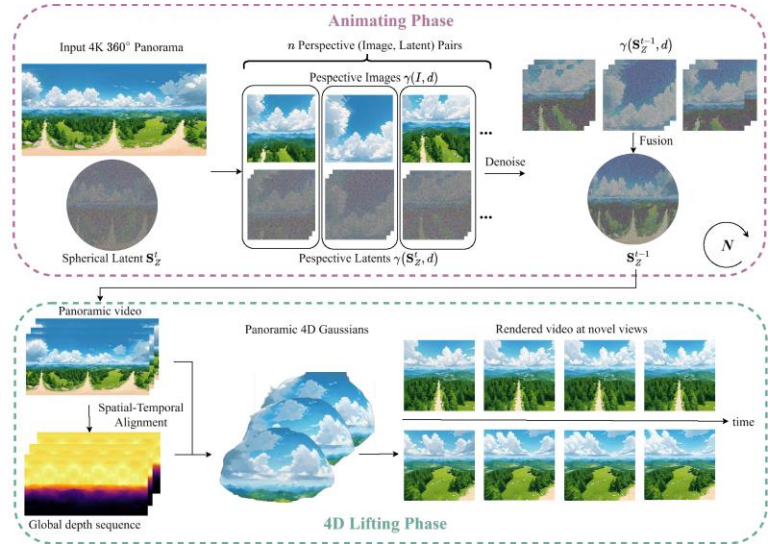
(iii) Creating 4D scenes that can be explored in VR devices, from the panoramic video by lifting it into dynamic 3D structures.



Highlight

4K4DGen takes a static panoramic image with a resolution of 4096×2048 and allows animation through user interaction or an input mask, transforming the static panorama into dynamic Gaussian Splatting. 4K4DGen supports the rendering of novel views at various timestamps, enriching immersive virtual exploration for VR devices.

TECHNIQUES



Omnidirectional Panoramic Representation

We represent the panoramic images or videos in the unit spherical space instead of the default equirectangular projected matrix. The relation between spherical representation $S(x,y,z)$ and the equirectangular projected image $I(u,v)$ is $S(x,y,z) = I(\frac{\operatorname{sgn} y}{\pi}\operatorname{arccot}\frac{x}{||y||}, \frac{2}{\pi}\arcsin z)$.

Animating from Panorama

We utilize a pre-trained generic diffusion model to consistently generate a panoramic video from a static panorama image. We use the pre-trained denoiser at each perspective view and fuse them into panoramic latent code at each denoising step.

4D Lifting from Panoramic Video

Given a dynamic panoramic video, we estimate the depth of each perspective view at each time via an off-the-shelf depth estimator and optimize them into consistent dynamic 3D geometry using semantic, spatial, and temporal losses.

RESULTS

Table 1: **Comparison with 3D-Cinemagraphy.** We compare our method with 3D-Cinemagraphy using rendered images from 4D representations. The IQ, IA, and VQ models represent the image quality scorer, image aesthetic scorer, and video quality scorer, respectively, within the Q-Align assessment framework. Our method, 4K4DGen, consistently achieves superior performance in both image and video quality across these metrics. Furthermore, 4K4DGen performs better in our user studies in terms of visual quality (Quality), motion amplitude (Amplitude), and the motion naturalness (Naturalness). Please refer to D.2 for further details.

Method	Q-Align (IQ) ↑	Q-Align (IA) ↑	Q-Align (VQ) ↑	Quality (UC) ↑	Amplitude (UC) ↑	Naturalness (UC) ↑
3D-Cinemagraphy (zoom-in)	0.47	0.38	0.57	7%	29.4%	19.7%
3D-Cinemagraphy (circle)	0.48	0.40	0.58	12%	32.0%	21.1%
Ours (holistic pipeline)	**0.66**	**0.44**	**0.62**	**81%**	**38.6%**	**59.2%**