A Implementation Details

A.1 Dual-branch Video Generation Model

Dataset. We use the Matterport3D dataset [5], which contains 10,800 panoramic images from 90 building-scale indoor scenes, for training the dual-branch image diffusion model. Text descriptions for both panoramic and perspective views are obtained using BLIP-2 [24] following [62]. Also, we adopt WEB360 [53], a high-quality panorama video dataset comprising 2,114 text-video pairs in 720P ERP format, with each video consisting of 100 frames. We uniformly sample 10 frames per video for image-level training. WEB360 is used for finetuning the dual-branch video model. Captions for perspective images and videos are generated using CogVLM [54].

Training Details. We use Stable Diffusion 1.5 [3] as the backbone for both branches, with rank-4 LoRA modules added to the spatial layers. Following [41, 62], we train the image diffusion model for 10 epochs using the AdamW optimizer with a batch size of 64 and a learning rate of 2e-4. After pretraining, we integrate the motion module from AnimateDiff [11] and bidirectional cross-attention modules, and train the video diffusion model on WEB360 for 10K steps using AdamW (batch size 8, learning rate 1e-4). We adopt DDIM sampling with 50 steps for inference. All training is performed on 8 NVIDIA A100 GPUs.

A.2 Geometry-aligned Reconstruction Model

Training Details. In 3DGS optimization, each panoramic frame is projected into $K = 20$ overlapping tangent perspective views to supervise Gaussian Splatting training. During projection, the camera's orientation is randomly initialized by uniformly sampling Euler angles across all three rotational axes (X, Y, Z) over the full 360-degree sphere. The camera's field of view (FOV) is set to 90-degree with perspective resolution fixed at 512 × 512 pixels. We extract extrinsic parameters from the estimated camera poses $\{P^t\}_{t=1}^{T}$ and use the intrinsic parameters to initialize scene cameras. Each panoramic frame is optimized for 15,000 epochs, employing both RGB loss and geometric correspondence loss throughout the entire training process. Additionally, the distillation semantic loss is applied between epochs 5,400 and 9,000. We set losses weights as $\lambda_1 = 0.8, \lambda_{ssim} = 0.2, \lambda_{lpips} = 0.05, \lambda_{sem} = 1, \lambda_{geo} = 0.05$. All training and experiments are performed on 4 NVIDIA 3090 GPUs.

B Loop Consistency Visualization

We stitch the left and right ends of the generated panoramic videos and visualize the results in Fig. 8 and Fig. 9. As shown, our method produces panoramas with smooth and seamless transitions at the boundaries, demonstrating strong consistency and continuity across the entire 360-degree view. **We provide more video demos in the supplementary materials.**

C Limitations and Future Work

In this work, we propose a novel text-to-dynamic panoramic 4D generation pipeline that integrates a dual-branch video generation model with a geometry-consistent reconstruction framework. While our method demonstrates strong performance across various benchmarks, there are still some limitations. First, our model is built upon pretrained image diffusion backbones, such as Stable Diffusion, to leverage their powerful generative capabilities. However, this reliance also imposes constraints, as the performance and flexibility of our approach are inherently limited by the capabilities and biases of the underlying pretrained models. Second, the scarcity of high-quality and diverse panoramic video datasets poses a challenge for generalization. While we make use of WEB360 and Matterport3D, these datasets still lack the variety and coverage to fully support broad, open-world scene generation and reconstruction. In future work, we plan to explore more robust and expressive base models to further enhance generation quality and flexibility. Also, we aim to curate or construct larger-scale, high-diversity panoramic video datasets to improve the generalization and adaptability of our pipeline.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

A **wood house** with a sloping roof sits beside a **calm lake** with autumnal foliage and a nearby **gravel pathway**.

Red wooden cabins with white trim, overlooking a calm, reflective body of **water**, under a starry **night sky** illuminated by the vibrant green **Northern Lights** arching over snow-capped **mountains**.

A picturesque **countryside** scene unfolds as rolling **green hills** stretch beneath a **blue sky**, revealing a winding **road** and charming **village** nestled amongst the fields.

Clear skies with wispy **clouds**, a reflective body of **water**, a **sandy shore** with limited vegetation, distant mountains under a vast horizon.

Figure 8: Loop Consistency Visualization.

Vivid **desert** landscape featuring sunlit **valleys** and rugged terrains.

Crystal-clear turquoise **waters**, pristine white **sandy shores**, a distant **island** with lush greenery and a few structures, and a vast expanse of the azure **sky** with wisps of **clouds**.

Waterfall cascades down rocky **cliffs**, surrounded by lush greenery and a serene **river**, under a vast **sky** with **scattered clouds**.

Sunset over the prairie, casting a golden glow on the grassy plains, with tufts of **wildflowers** dotting the landscape and a few **clouds** drifting in the vast **blue sky**.

Figure 9: Loop Consistency Visualization.