Matrix-3D

OMNIDIRECTIONAL EXPLORABLE 3D WORLD GENERATION

Image / Text Input

Generated 3D world
*freely explored from **any position** and in **any direction***

Agenda

- Motivation, Contribution, Related Work
- Framework
- Text to Panorama Generation
- Trajectory Guided Panorama Video Generation
- 3D World Generation
- Matrix-Pano Dataset
- Comparison
- Ablation Study

Motivation

Spatial intelligence is a rapidly developing technology that enables machines to model the 3D world and to understand, analyze, and reason about it. The modeling and generation of 3D world are key modules of spatial intelligence.

High-quality, diverse 3D scenes can serve as virtual environments for AI training and testing, enhancing AI's generalization and adaptability in the field of autonomous driving [64] and embodied intelligence [10], thus advancing the development of world models.

Additionally, 3D scene generation has widespread applications in game design [24], film production [19], and virtual reality [41].

Contributions

Matrix-Pano Dataset

a scalable and high-quality panoramic video dataset with precise camera poses, depth maps, and text annotations, tailored for trajectory-guided panoramic video generation and wide-coverage 3D world reconstruction

Novel trajectory-guided video diffusion model by scene mesh renders

alleviates Moiré patterns and incorrect occlusion relationships, leading to improved visual quality in generation

Panoramic 3D reconstruction methods

two types of panoramic 3D reconstruction methods to achieve rapid and detailed 3D reconstruction, respectively

Related Work

3D World Generative Models

1. Diffusion model generates novel views conditioned on sparse or singleview inputs and target camera poses;
2. Per-scene optimization is performed based on the generated views and their associated poses.

Focus on perspective image inputs, which limits their ability to recover omnidirectional 3D structures

Camera-controlled Video Diffusion Models

These methods utilize different forms of camera conditioning, including camera extrinsic parameters, Plücker embeddings, and point cloud renders.

Fails to provide precise camera control. Point cloud renders offer improved controllability, but they frequently suffer from Moiré patterns

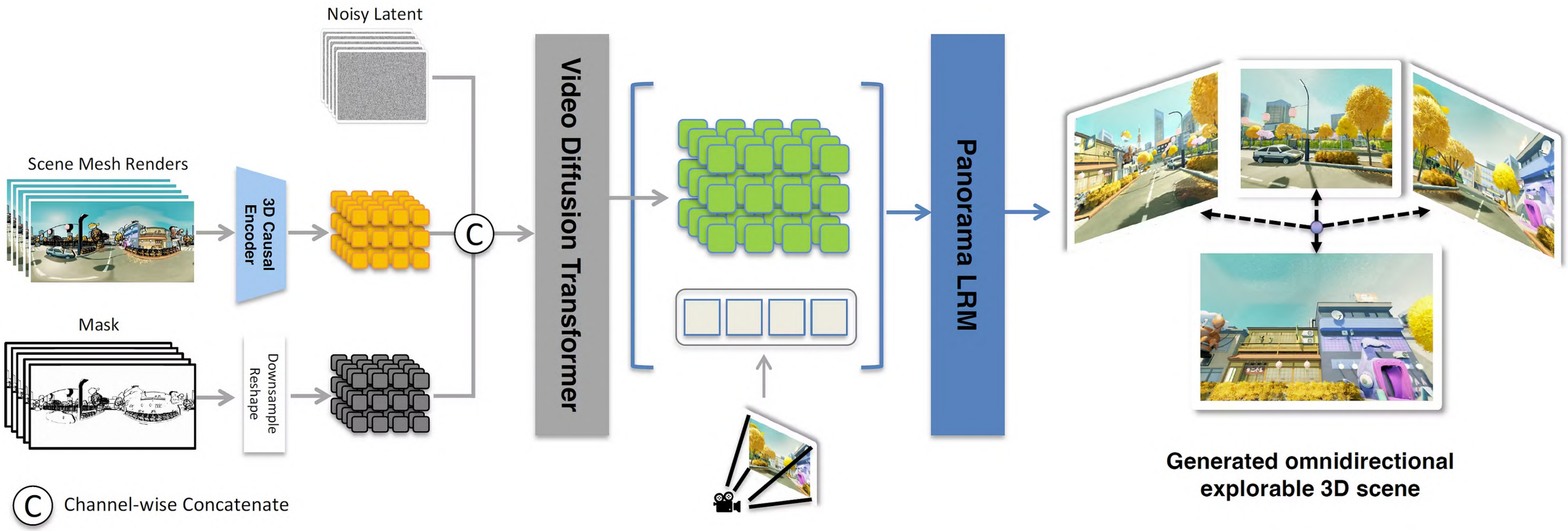
Panorama Generation Models

Diffusion-based panoramic image generation models have achieved notable progress. Inherently lack information about physically occluded regions.

360DVD introduced the WEB360 dataset, establishing an early benchmark for text-to-panoramic video.

None of these methods support the generation of panorama videos with precise scene trajectory control or the conversion of panorama videos into 3D worlds.

Framework



Given trajectory guidance in the form of scene mesh renderings and corresponding masks,
obtained by rendering an estimated mesh along a user-defined camera trajectory,
we train an image-to-video diffusion model to generate high-quality panoramic videos that precisely follow the specified trajectory.
The generated 2D panoramic content is then lifted into an omnidirectional, explorable 3D world
using a large-scale panorama reconstruction model..

Custom panorama rendering pipeline built on Movie Render Queue (Unreal)

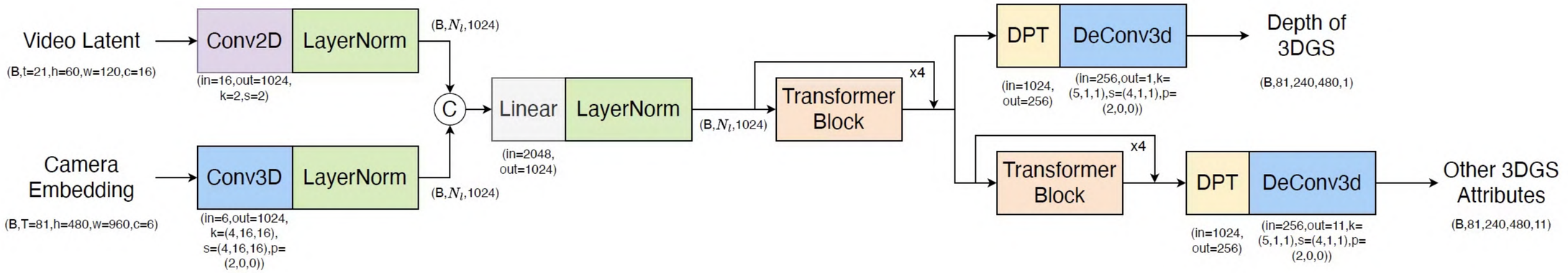


Figure 12: The network architecture of our large panorama reconstruction model.

For text-to-panorama generation, we train a LoRA on FLUX with 1000 selected panorama images from the Matrix-Pano dataset.

Our image-to-panorama generation pipeline builds directly upon WorldGen.

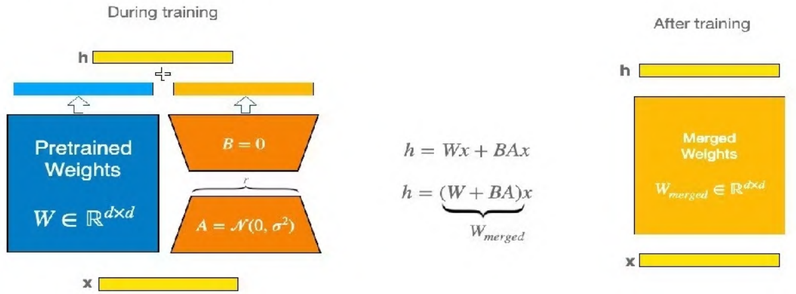
During the inference stage, we apply latent rotation and circular padding

to enable the generation of loop-consistent results following PanFusion

LoRA

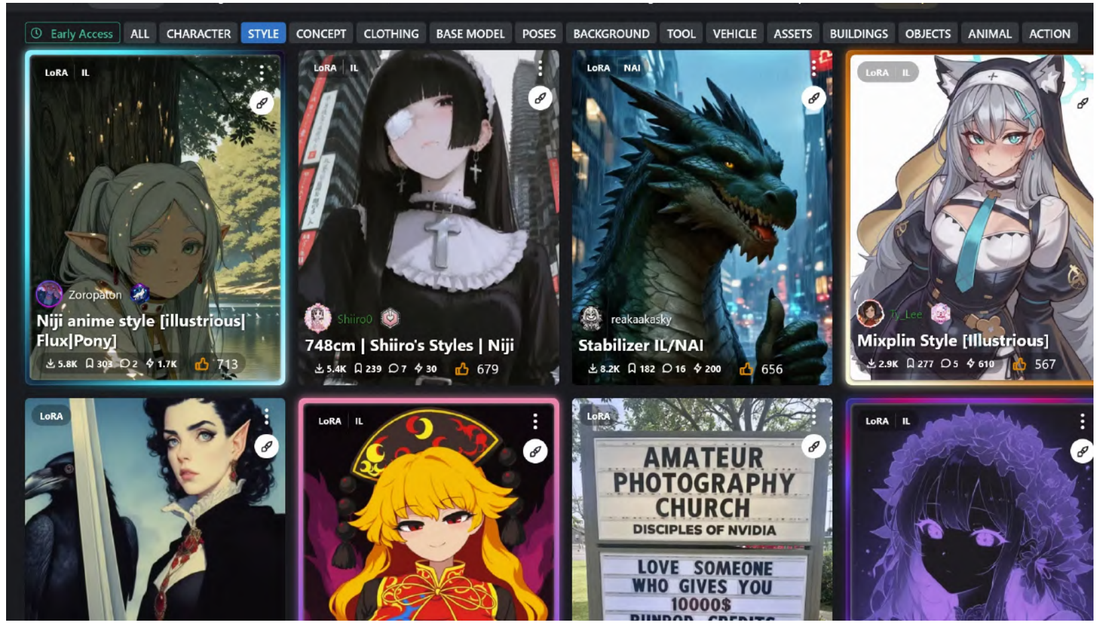
LoRA: Low-Rank Adaptation of Large Language Models

Edward Hu* **Yelong Shen*** **Phillip Wallis** **Zeyuan Allen-Zhu**
Yuanzhi Li **Shean Wang** **Lu Wang** **Weizhu Chen**
Microsoft Corporation
{edwardhu, yeshe, phwallis, zeyuana,
yuanzhil, swang, luw, wzchen}@microsoft.com
yuanzhil@andrew.cmu.edu



$$h = Wx + BAx$$

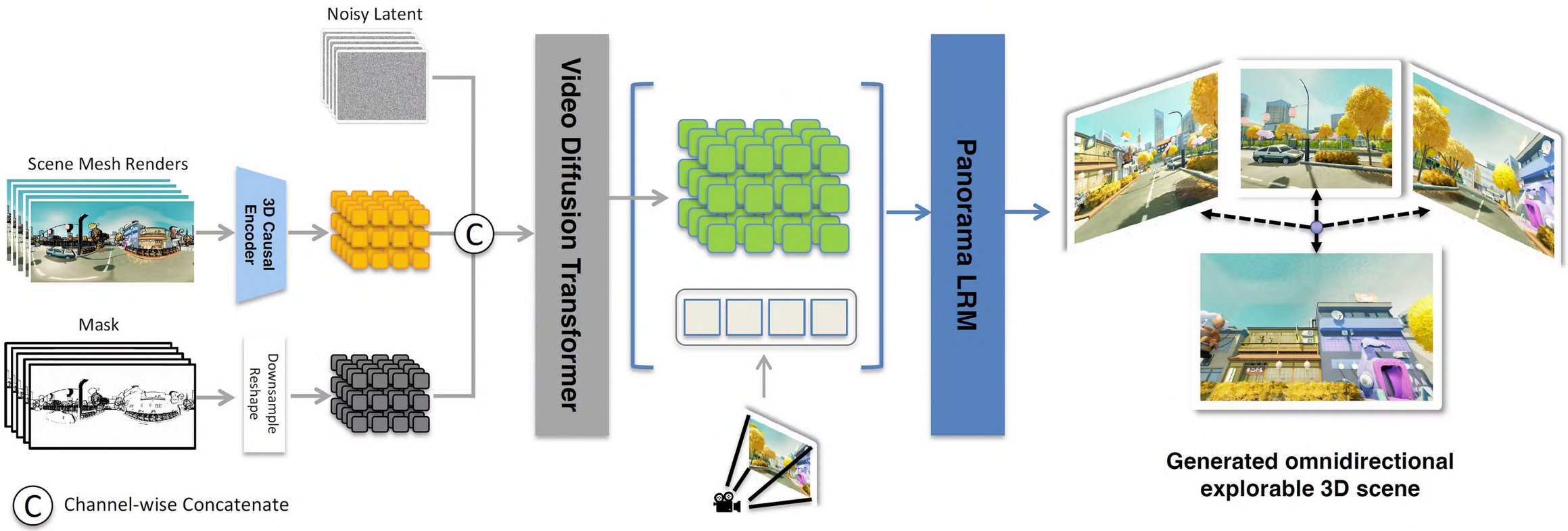
$$h = \underbrace{(W + BA)}_{W_{merged}} x$$



LoRA models

Agenda

- ~~Motivation, Contribution, Related Work~~
- ~~Framework~~
- ~~Text to Panorama Generation~~
- Trajectory Guided Panorama Video Generation
- 3D World Generation
- Matrix-Pano Dataset
- Comparison
- Ablation Study



Given trajectory guidance in the form of scene mesh renderings and corresponding masks,
obtained by rendering an estimated mesh along a user-defined camera trajectory,
we train an image-to-video diffusion model to generate high-quality panoramic videos that precisely follow the specified trajectory.
The generated 2D panoramic content is then lifted into an omnidirectional, explorable 3D world
using a large-scale panorama reconstruction model..

Trajectory Guided Panorama Video Generation

Trajectory Guidance Construction

Given a panorama with its depth map, we construct the initial scene mesh by projecting depth values into world coordinates to get mesh vertex positions.

Then these vertex are connected to mesh faces according to their connectivity in the pixel space. The color of each mesh vertex is determined by its corresponding pixel color.

To represent invisible areas of and accurately capture its occlusion relationship with visible parts, we further select pixels with drastically changing depth, mark their corresponding vertices as invisible and assign pure black color on them.

In detail, the depth change of each pixel is calculated the as depth variation among its 1-ring neighbor pixels, and those pixels with depth change larger than a predefined threshold will be marked as invisible.

Panoramic Video Generation

Given the trajectory-guided video sequence, encode the video using the video model's 3D causal VAE encoder to obtain video latents.

Corresponding mask sequence is downsampled and reshaped to mask latents that match the dimension of the video latents.

Video and mask latents are concatenated with the noisy latent and then fed into the Diffusion Transformer for denoising.

Global semantic guidance: inject the input panorama with its correlated or annotated text prompt into the network.

Initial image is encoded to obtain an image-level embedding, which is then fused with the text prompt embedding and integrated into the generation process via cross-attention.

Trajectory Guidance Construction

Initial Panorama with depth

point cloud

mesh

3D World Generation

Optimization-based 3D Reconstruction

The optimization-based method takes the generated video with the input camera trajectory as input.

Construct a set of keyframes: one panoramic frame every five frames, and use only these keyframes for optimization. Estimate the panorama depth of the keyframes with MoGe, perform the least squares registration to align these resulting depth maps, and use their corresponding world-coordinate point clouds as the initialization for 3DGS blobs.

Original 3DGS optimization pipeline only accepts perspective images as input, we first crop each panoramic keyframe into 12 perspective images. Apply StableSR] to perform super-resolution on these perspective images, which are subsequently fed into the 3DGS optimization pipeline.

The loss between the rendered images and the input images is used as the objective function.

Large Panorama Reconstruction Model

Given the video latent representation and the corresponding camera poses encoded as *spherical Plücker embeddings*, we first transform them into latent tokens and pose tokens using dedicated patchify modules to ensure equal sequence lengths. These tokens are then concatenated along the channel dimension and fed into a series of base Transformer blocks. For Gaussian attribute prediction, we adopt the DPT head. Since the DPT head performs upsampling only along the spatial dimensions, we further employ a 3D deconvolution layer to upsample along the temporal dimension, ensuring alignment with the original video sequence on temporal dimension.

The 3D deconvolution layer produces the 3D Gaussian attributes as a 12-channel tensor

Large Panorama Reconstruction Model

Optimization

Straightforward approach for optimizing our model: involve rendering panoramic images for loss computation, we empirically found that using panoramic images as direct supervision introduces sparsity artifacts when rendered with a perspective 3DGS rasterizer.

Alternative strategy: for each selected panoramic image, we first generate 12 perspective-view patches that collectively cover the full 360◦ field.

Because using all 12 patches as supervision is computationally prohibitive, we randomly sample one patch per panorama and render it with a standard perspective rasterizer.

Augment this reference set with both interpolated and extrapolated views. During training, we then compute our reconstruction loss between each rendered perspective patch and its corresponding ground-truth crop, driving accurate panoramic 3D reconstruction.

Two Stage Training Strategy

To overcome the challenge of panoramic depth estimation from video latent, we adopt a two stage training strategy.

First stage: initialize model training with depth loss.

Second stage: optimize the remaining GS attributes using an image reconstruction loss that combines mean squared error (MSE) and perceptual similarity (LPIPS) loss.

Two Stage Training Strategy

First stage

To overcome the challenge of panoramic depth estimation from video latent, we adopt a two stage training strategy.

First stage: initialize model training with depth loss.

Goal: predict depth for each frame based on the video latent.

Second Stage

Second stage: optimize the remaining GS attributes using an image reconstruction loss that combines mean squared error (MSE) and perceptual similarity (LPIPS) loss.

Freeze the depth-prediction parameters and update only the remaining Gaussian attributes.

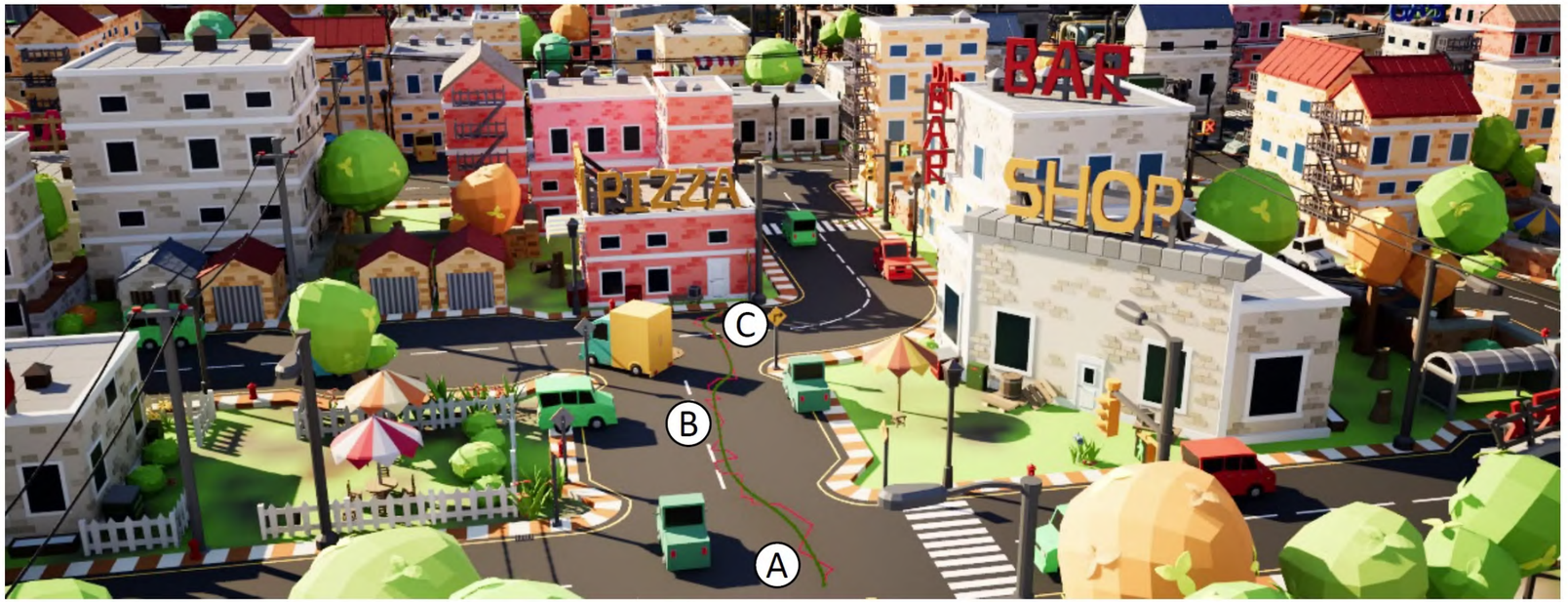
Avoid overfitting to observed views and improve generalization to novel viewpoints: randomly sample 32 reference views comprising three categories: views seen in the context frames, interpolated views, and extrapolated views. Each panoramic view is then cropped into 12 perspective images that can include the whole panorama at 512 × 512 resolution, with a randomly selected field of view between 60◦ and 120◦.

These crops serve as supervision signals during GS attribute fine-tuning

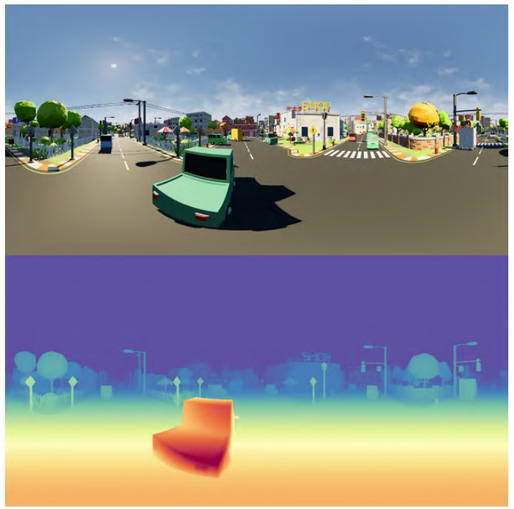
Agenda

- ~~Motivation, Contribution, Related Work~~
- ~~Framework~~
- ~~Text to Panorama Generation~~
- ~~Trajectory Guided Panorama Video Generation~~
- ~~3D World Generation~~
- Matrix-Pano Dataset
- Comparison
- Ablation Study

Matrix-Pano Dataset



Ⓐ Ⓑ Ⓒ



Matrix-Pano Dataset

Matrix-Pano Dataset, a scalable synthetic dataset designed for generating omnidirectional explorable 3D worlds. The construction process of consists of the following steps.

Step 1: Data Collection in Unreal Engine. Unreal Engine allows for flexible data preparation across a variety of scenes, perspectives, and content. Utilizing 504 high-quality 3D scenes, we encompass a wide range of indoor and outdoor settings.

Step 2: Exploration Route Sampling. We developed a trajectory sampling algorithm to generate plausible and visually coherent camera paths. For each scene, we first identify walkable surfaces, e.g., roads or floors, then apply the Delaunay triangulation algorithm, which creates a set of nonoverlapping triangular meshes from sparse points on a two-dimensional plane.

Matrix-Pano Dataset

Step 3: Collision Detection. We implement a collision detection mechanism to remove trajectories that cause geometry clipping or object intersections, which can degrade quality and stability. Using a bounding box proxy algorithm, objects are simplified into 3D bounding boxes based on their nearest and farthest points, balancing spatial accuracy with computational efficiency. Trajectories are simulated step-by-step, and any intersecting paths are discarded.

Step 4: Data Annotation and Quality Filtering. We ensure the quality of the dataset through two filtering stages:
(1) Automatic Filtering: We use Video-LLaMA3 [65] to evaluate videos based on detailed quality, semantic information, and motion richness, filtering out low-quality content.
(2) Manual Assessment: The first frame of each video is manually reviewed to remove samples with poor rendering quality or missing details. Video-LLaMA3 automatically annotates videos to support text-controlled and multimodal tasks.

Agenda

- ~~Motivation, Contribution, Related Work~~
- ~~Framework~~
- ~~Text to Panorama Generation~~
- ~~Trajectory Guided Panorama Video Generation~~
- ~~3D World Generation~~
- ~~Matrix-Pano Dataset~~
- Comparison
- Ablation Study

Comparison with Panorama Video Generation Model

Table 2: Comparison of Panoramic Video Generation and Camera Guided Generation Models.

Models	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	FVD ↓	R_{err} ↓	T_{err} ↓
Panoramic Video Generation Models							
360DVD [53]	9.65	0.349	0.834	112	2700	–	–
Imagine360 [46]	11.6	0.391	0.599	66.7	1600	–	–
GenEx [33]	16.1	0.600	0.380	42.2	1110	–	–
Matrix-3D 480p	23.7	0.722	**0.0776**	15.4	234	–	–
Matrix-3D 720p	**23.9**	**0.747**	0.0907	**11.3**	**140**	–	–
Camera Guided Generation Models							
ViewCrafter [63]	21.6	0.701	0.161	47.3	762	0.0940	0.0453
TrajectoryCrafter [62]	21.8	0.682	0.126	33.1	675	0.0338	0.0488
Matrix-3D 480p Persp.	24.1	0.750	0.113	23.9	438	0.0325	0.0310
Matrix-3D 720p Persp.	**24.3**	**0.777**	**0.108**	**12.5**	**165**	**0.0306**	**0.0297**

ODGS | Ours (Feed-forward) | Ours (Optimization) | Ground Truth

Comparison with 3D World Reconstruction

3: Quantitative comparison of 3D world reconstruction on the benchmark dataset. All metrics t for time are calculated on perspective images cropped from the panorama renderings. Our ization-based pipeline achieves the best performance in terms of visual quality, while the forward pipeline enables much faster reconstruction.

Methods	PSNR ↑	LPIPS ↓	SSIM ↑	Time (↓) [s]
ODGS [23]	22.04	0.444	0.673	745
Ours (Feed-forward)	22.30	0.389	0.647	**10**
Ours (Optimization-based)	**27.62**	**0.294**	**0.816**	571

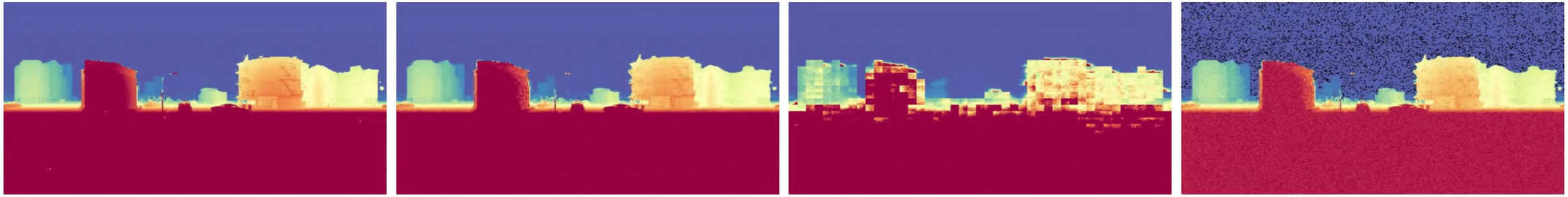
Comparison with 3D World Reconstruction

Ablation Study

Initial Panorama | Point cloud renders condition | Scene mesh renders condition | Ground Truth

Rendered video from mesh

Ablation Study



GT depth | Depth by DPT head | Depth by Deconv3D head | Depth after finetuning

Figure 11: Comparison between depth predictions from the DPT head and the 3D deconvolution head. When depth-related parameters are not frozen, depth degrades during second-stage training.

The effectiveness of DPT head

Conclusions and Future Work

Conclusions

- Introduce a trajectory-guided panoramic video diffusion model conditioned on mesh renders, which produces visually coherent and structurally faithful scene videos.
- Lift the generated videos into 3D: two separate reconstruction pipelines: a feed-forward model for fast scene reconstruction and an optimization-based approach for high-fidelity geometry recovery.

Limitation

- 3D scene generation model is built upon a video diffusion model, its inference speed is relatively slow, with the generation of a single scene taking tens of minutes.
- Matrix-Pano dataset, unnatural transitions in depth values occasionally occur in semi-transparent or perforated regions, such as those found in trees and fences.
- Estimating depth from the video latent representation is particularly challenging: the latent compresses the original video and encodes only appearance cues, since the video VAE’s objectives do not incorporate geometric information.

Conclusions and Future Work

Future Work

- 3D scenes generated by Matrix-3D typically include information only from currently visible areas. Future research should investigate methods to generate scenes for unseen areas, such as by employing specific trajectory settings or integrating 3D object generation into the existing pipeline.
- Editability of the generated 3D worlds can be further enhanced. This involves enabling user-driven operations such as scene modification and semantic-level interactions, such as issuing high-level commands like “add a tree beside the house” or “remove the car from the road”. Enhancing editability would make the system more adaptable for downstream applications in digital content creation, simulation environments, and embodied AI systems.
- Extend our method to dynamic scene generation, enabling each object in the scene to move and interact, thereby providing users with a more immersive experience and further advancing research in world models.