PanoDreamer: Consistent Text to 360-Degree Scene Generation

Zhexiao Xiong[1,2*]   Zhang Chen[1]   Zhong Li[1]   Yi Xu[1]   Nathan Jacobs[2]

[1] OPPO US Research Center   [2] Washington University in St. Louis

Abstract

Automatically generating a complete 3D scene from a text description, a reference image, or both has significant applications in fields like virtual reality and gaming. However, current methods often generate low-quality textures and inconsistent 3D structures. This is especially true when extrapolating significantly beyond the field of view of the reference image. To address these challenges, we propose PanoDreamer, a novel framework for consistent, 3D scene generation with flexible text and image control. Our approach employs a large language model and a warp-refine pipeline, first generating an initial set of images and then compositing them into a 360-degree panorama. This panorama is then lifted into 3D to form an initial point cloud. We then use several approaches to generate additional images, from different viewpoints, that are consistent with the initial point cloud and expand/refine the initial point cloud. Given the resulting set of images, we utilize 3D Gaussian Splatting to create the final 3D scene, which can then be rendered from different viewpoints. Experiments demonstrate the effectiveness of PanoDreamer in generating high-quality, geometrically consistent 3D scenes.

1. Introduction

The immense potential of text-to-3D applications in VR/AR platforms, industrial design, and the gaming industry has driven substantial research efforts toward establishing a robust approach for immersive scene content creation. Recent developments in diffusion models [12, 25, 46] make it possible to generate high-quality, geometrically-correct images from text, allowing for customized 2D content generation.

Based on the recent advance in 2D text-to-image generation [21, 22, 41, 50], many works have begun focusing on 3D scene generation. Some works [5, 6, 44] first generate an initial point cloud based on a reference image, employing a progressive warp-and-refine approach to complete the 3D scene reconstruction. However, due to the limited camera field-of-view (FoV), these approaches require multiple

iterations to generate a complete scene, with each iteration relying solely on information from the previous stage. As a result, error accumulation from monocular depth estimation and artifacts from diffusion generation hinder these models' ability to maintain long-term geometric and appearance consistency, particularly with large camera movements.

To overcome these challenges, recent works have leveraged panorama-to-3D scene generation [30, 52] to generate scenes with a larger FoV. Utilizing advancements in text-to-panorama generation [47], these methods use panoramas as intermediate representations of the 3D scene, subsequently obtaining 3D representations using Neural Radiance Fields (NeRF) or 3D Gaussian Splatting (3D-GS). However, since the geometry is based on a single panorama, the generated 3D scenes have a limited spatial extent and are significantly impacted by occlusions. As a result, users have limited freedom to move about the scene, greatly limiting the usefulness of the 3D model.

In this work, we propose PanoDreamer, a novel framework that enables global-level scene generation with geometric consistency and allows for customized 3D scene extension. Our approach adopts a multi-stage pipeline: first generating a static panoramic scene, followed by extending the scene dynamically based on user-defined initial images and camera trajectories. To generate the static panoramic scene, given a text prompt and/or a user-provided reference image, we synthesize images from an initial viewpoint using an LLM engine and composite them into a complete equirectangular panorama. This panorama is then lifted into 3D to create an initial point cloud. We then generate a set of additional images from different viewpoints. We use a view-conditioned video diffusion model to generate sequences based on user-specified initial images and trajectories, enabling both continuous, geometrically consistent scene generation and flexible control over viewpoint shifts.

The resulting point clouds are composed into a global point cloud using depth alignment, followed by 3D Gaussian Splatting to produce the 3D scene representation. To enhance the scene completeness, we propose a strategy to generate a set of supplementary views and employ a semantic-preserving generative warping framework to inpaint occluded regions. These supplementary viewpoints,

---

along with their inpainted images, are used to refine the 3D Gaussians, thereby reducing artifacts and enhancing scene completeness.

Our main contributions can be summarized as follows:

- We propose PanoDreamer, a holistic text to 360-degree scene generation pipeline, which achieves consistent text-to-360-degree scene generation with customized trajectory-guided scene extension.
- We introduce semantically guided novel view synthesis into the refinement of 3D-GS optimization, reducing artifacts and improving geometric consistency.
- Experiments show the effectiveness of our model in generating geometrically consistent and high-quality 360-degree scenes.

2. Related Work

Panorama Generation With the development of diffusion models [24], many studies have sought to generate panoramic scenes using existing text-to-image diffusion techniques [1, 2, 29]. These methods often use text-to-image generation models or image outpainting techniques to first synthesize multi-view images, subsequently generating the panorama through equirectangular projection. MVDiffusion [29] proposed a correspondence-aware attention to generate text-conditioned multi-view images or extrapolates one perspective image to a full 360-degree view. Some later methods finetune the diffusion models to generate panoramas. StitchDiffusion [31] used Low-Rank Adaptation (LORA) [10] to generate panoramic images and achieves customized generation, PanFusion [48] tried to use panoramic diffusion in the latent space, Diffusion360 [8] utilized DreamBooth [25] finetuning alongside circular blending to produce panoramas in both text-to-panorama and single-image-to-panorama tasks, and MVPS [40] used geospatial information to guide panorama generation. Our method supports both text-only input and combined text and image conditions, achieving flexible panorama generation.

Conditional Video Diffusion Models With the increasing need for multi-modality control, controlled video generation have evolve rapidly. Benefiting from previous customized image generation methods [13, 41, 50], conditional video diffusion models also allow for multiple control guidance, including text [3, 4, 27, 39], RGB images, depth [7], semantic maps [20] and trajectory [19, 43]. Recent studies regard video diffusion models as a strong tool in downstream tasks, such as stylization [14], motion control [34], novel view synthesis [45]. Specifically, view conditioned video diffusion models such as ViewCrafter [45] regarded point cloud renders as control to synthesis novel view generic scenes either from both single or sparse images, which enables a point cloud completion and benefit the downstream tasks. Our work leverage ViewCrafter's

generalization ability to help customize the extention of scene generation with geometric consistency.

Dreaming-based 3D Scene Generation Synsin [36] was one of the pioneering methods that employed a warp-and-refine strategy to generate point clouds of a scene. With the rapid advancement of diffusion models and Score Distillation Sampling (SDS)-based techniques, dreaming-based text-to-3D generation has emerged as a popular approach for creating 3D content. Early methods predominantly utilized Neural Radiance Fields (NeRF)[15, 49] or mesh-based representations[9, 28] to achieve scene reconstructions. More recent works, such as LucidDreamer [6], have employed 3D Gaussian Splatting [11] to achieve consistent rendering with greater geometric flexibility. However, these approaches primarily focus on generating forward-facing scenes, which restricts the scalability for more extensive scene generation.

To achieve global-level 360-degree scene generation, some recent works have turned to using panoramic representations as an intermediate stage for comprehensive scene synthesis. PERF [30] was the first to propose a 360-degree novel view synthesis method by training a panoramic neural radiance field from a single panorama, enabling free movement within a 3D environment. Despite its potential, this approach remains largely confined to indoor scenes. Later methods like DreamScene360 [52] and Holo-Dreamer [51] advanced the concept by adopting panoramic Gaussian splatting for 360-degree scene generation. Additionally, approaches such as LayerPano3D [42] introduced layered panorama generation techniques to manage complex scenes, subsequently lifting these layers into 3D Gaussian splatting representations.

In our work, we propose a method for text-to-360-degree scene generation that also utilizes panoramas as an intermediate representation, combined with 3D Gaussian splatting to generate the final 3D scene. Our approach not only overcomes the limitations of previous methods but also enhances geometric consistency and provides greater flexibility in scene extension and customization.

3. Preliminary

3D Gaussian Splatting. 3D Gaussian Splatting (3DGS) [11] is a recent pioneer method for novel view synthesis and 3D reconstruction,utilizing the multiview calibrated images from Structure-from-Motion. Unlike implicit representation methods like NeRF [16], 3D-GS renders in an explicit manner through splatting, achieving real-time rendering and reduced memory consumption. The 3D Gaussians can be queried as:

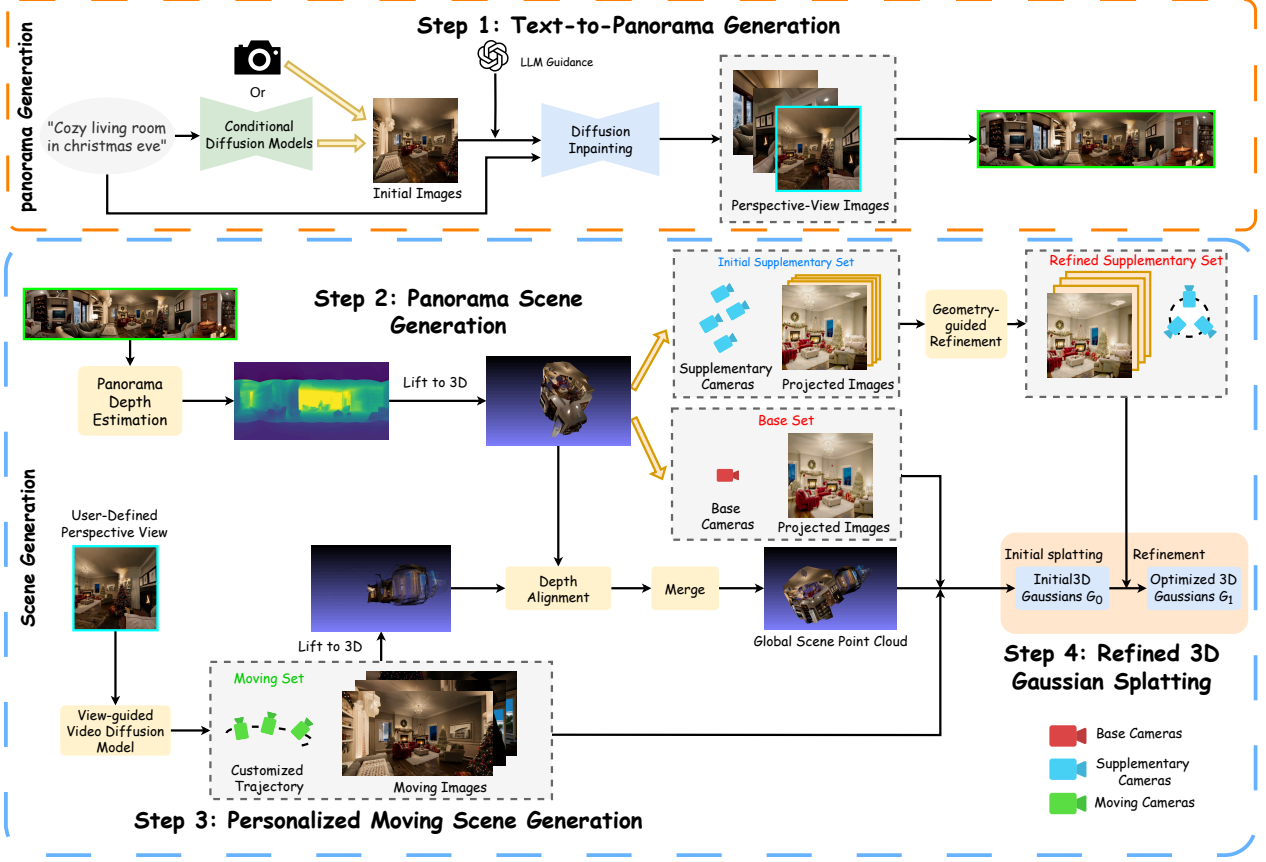$$G(x) = e^{-\frac{1}{2}(x)^T \Sigma^{-1}(x)}, \tag{1}$$

Figure 1. Modules of our proposed framework. (a) Text-to-Panorama Generation: we use LLM as guidance to guide the generation of perspective-view images. (b) Scene Generation: we divide it into static panorama scene generation and customized moving scene generation. Besides base camera set, we compose an additional supplementary camera set for the static panorama scene and use semantic-preserved warping to generate the missing region, which is used for 3d gaussian splatting refinement.

where x represents the distance between the center position μ and the query point. During the rendering process, the color of r on the image plane is rendered sequentially with point-based volume rendering technique through:

$$C(r) = \sum_{i \in \mathcal{N}} c_i \sigma_i \prod_{j=1}^{i-1} (1 - \sigma_j), \quad \sigma_i = \alpha_i G(x_i), \quad (2)$$

where $\mathcal{N}$ represents the number of sample points on the ray r, c_i and α_i denote the color and opacity of the i-th Gaussian, and x_i is the distance between the point and the i-th Gaussian.

4. Method

Given a text prompt T or an optional user-provided inintial image I, we aim to get 3D scene-representation with global-level consistency, and allows for free camera movements among different sub-scenes. In this section, we first introduce the static text-to 360-degree panorama generation in Sec. 4.1, then introduce the panorama scene generation

and supplementary moving scene generation in Sec. 4.2 and Sec. 4.3 respectively. Finally we introduce the scene generation with 2-stage 3D gaussian splatting in Sec. 4.4.

4.1. Text-to-Panorama Generation

Traditional text-to-panorama generation methods often rely on a single prompt, restricting their ability to generate panorama with rich content and may lead to repeated content. Inspired by L-Magic [2], we decompose the panorama generation process into two distinct stages: LLM-guided perspective image generation and panorama composition. First, an image is projected into the unit sphere by defining the vertices V on each image pixel and creating edges between adjacent pixels. Then we use a warp-and-inpainting strategy to get novel view images. Specifically, given a rotation matrix R from viewpoint A to viewpoint B, the pixel in the other frame is computed as $\mathbf{P}_{\text{rot}}^{\text{i}} = \mathbf{R}\mathbf{P}_0$, where the camerea field of view(FoV) is set to be 100 degrees. In this process, binary mask M is used to ensure that the inpainting is constrained in the non-overlapping region between adja-

cent views. In the inpainting stage, we use Stable Diffusion V2 inpainting model [24] to extrapolate the large missing region of the warped view. To effectively remove the artifacts in the inpainting stage, we use use GPT4-o to achieve instruction-guided inpainting, which helps remove the duplicated objects.

To remove the blurry region in the border of the perspective view images and enhance the detail, we use diffusion-based super-resolution [35] to increase the resolution of each perspective-view image from 512×512 to 2048×2048. We warp the high-resolution image to a lower-resolution next-view image and use super-resolution again. This iterative process is repeated to obtain high-resolution images for all perspective views.

Finally, equirectangular projection is used to warp all perspective views to the same equirectangular plane and merge into a seamless panorama. Since the merged panorama does not cover the full 180-degree FoV in the vertical direction, we utilize a panorama inpainting method [38] to fill in the missing regions at the top and bottom. Eventually, we can get a complete equirectangular panorama with 180-degree vertical FoV.

4.2. Panorama Scene Generation

After synthesizing the panorama, we use a zero-shot panorama depth estimation network [32] to get the depth of the panorama, based on which we lift the panorama to 3D and get the point cloud of the panorama. As the lifted point cloud only contains point clouds generated from the same location, artifacts exists due to the occlusions. In the boundary area of the objects, the point cloud is not continuous, leading to holes in the projected image when translating camera positions. Therefore, we propose a diffusion-based refine method to refine the projected images.

Previous scene generation approaches often employed a warp-and-refine strategy to progressively generate the scene. However, this strategy heavily relies on the accuracy of monocular depth estimation, often leading to blurry results in complex scenes with noisy depth maps. Moreover, semantic details are frequently lost, particularly during large viewpoint changes. A recent work, GenWarp [26], has demonstrated effective geometric and semantic-preserved warping by augmenting cross-view attention with self-attention in the diffusion process. Based on this, we use a geometry-preserved warping method to fill in missing points in the point cloud. Specifically, we first construct a base camera set $P_B = p_1, p_2, \cdots, p_m$, where each camera is positioned at the center of the point cloud and uniformly faces different directions across 360-degree. We then project images from these cameras to form the base image set $I_B = I_1, I_2, \cdots, I_m$. Since the cameras are positioned at the center of the point cloud, these base images are largely free from artifacts. Subsequently, we sample an

additional $4m$ supplementary cameras C_S, all sharing the same intrinsics K. For each base camera c_m, we translate it up, down, left, and right by a uniform offset, forming a supplementary camera set $P_S = p_{(m,1)}, p_{(m,2)}, p_{(m,3)}, p_{(m,4)}$.

Based on the base image I_m and the corresponding depth map D_m obtained via multiview depth estimation with Open3D, we conduct the semantic-preserved generative warping in the latent space through GenWarp, represented as:

$$E_{m,n} = \text{GenWarp}\left(E_m; D_m, P_{m \rightarrow (m,n)}, K\right), \quad (3)$$

where E_m and $E_{m,n}$ represent the Fourier features of the base image and the projected image from the supplementary camera respectively, $P_{m \rightarrow (m,n)}$ denotes the relative camera pose from the base camera to the corresponding supplementary camera. The resulting feature embedding is then decoded to yield the inpainted supplementary image $I_{m,n}$. We also get the occlusion mask $M_{m,n}$, represented as the supplementary mask set M_S, which is also shown in Fig. 1. Therefore, we get the supplementary camera set P_S and the corresponding supplementary image set I_S based on semantic-preserved generative warping. We save I_B, P_B, I_S, P_S and M_S for the subsequent 3D-GS optimization.

4.3. Supplementary Moving Scene Generation

Although the panorama point cloud has a 360-degree view of the scene, it is often limited to a single position and restricted by occlusions such as walls and furniture, especially for indoor scenes like a single room. In practical applications, users may expect to navigate to other areas to obtain more comprehensive, global views of the environment. To enable this, we utilize view-guided video diffusion models [45] to generate moving scenes as supplementary views. Users can select a perspective-view image generated during the text-to-panorama process as the initial reference and define a custom camera trajectory.

Specifically, based on the initial image and the user-defined camera trajectory, we render a sequence of video frames along a target direction outside the panorama scene. Leveraging the reconstruction capability of the latest dense stereo methods, such as DUSt3R [33], we lift the moving scenes into a 3D point cloud, while also capturing the camera poses for each frame in the world coordinate system. However, directly aligning the moving scenes with the panorama scene results in two challenges: (1) overlapping regions between scenes, and (2) depth misalignment due to differences in the depth estimation methods used for the panorama and the moving scenes.

To mitigate these issues, we implement a masking strategy that eliminates duplicate regions by excluding pixels corresponding to the initial frame from the point cloud. Given the initial image I_0 and a specified trajectory, we generate m frames of images $I_1, I_2, \cdots, I_m$ using view-guided

Figure 2. Semantic-preserved Refinement: For each base camera, we apply supplementary cameras to up, down, left, and right directions respectively. For each supplementary camera, we get projected images through semantic-preserved generative warping [26] to fill the missing area brought by occlusion.

video diffusion models. We then uniformly sample n images from the video sequence and perform sparse-view reconstruction with DUSt3R.

In the selected set of images, the first frame overlaps with the initial image, and we ensure that all the camera poses of the frames are in the same world coordinate system. To address the overlapping regions between the scenes, we create a view mask M_0 for the points that lie within the view of the first camera. We first transfer the point cloud from the world coordinate to the camera coordinate, represented as:

$$\mathbf{P}_{cam} = \mathbf{T}_{w2c} \cdot \mathbf{P}_h^T, \tag{4}$$

where $\mathbf{P}_{cam} \in \mathbb{R}^{512 \times 1024 \times 4}$. For the point cloud in the camera coordinate system, we check if the point is in the camera view through $M_{front} = (\mathbf{P}_{cam}[\ldots, 2] > 0)$, Next, we project the 3D points in the camera frame onto the image plane using the camera's intrinsic matrix $\mathbf{K}$ through:

$$\mathbf{u}_h = \mathbf{K} \cdot \mathbf{P}_{cam}[\ldots, :3]^T, \tag{5}$$

which gives us the homogeneous coordinates in the image plane. We then normalize these to obtain the pixel coordinates through:

$$\mathbf{u} = \frac{\mathbf{u}_h[\ldots, 0]}{\mathbf{u}_h[\ldots, 2]}, \quad \mathbf{v} = \frac{\mathbf{u}_h[\ldots, 1]}{\mathbf{u}_h[\ldots, 2]}, \tag{6}$$

where $\mathbf{u}$ and $\mathbf{v}$ are the horizontal and vertical pixel coordinates on the image plane, indicating the column and row position of the 3D point when projected onto the camera image, respectively. The overall mask $\mathbf{M}$ is then computed as:

$$\mathbf{M} = \neg(\mathbf{M}_{bound} \cap \mathbf{M}_{front}), \tag{7}$$

which indicates the indices of the points that are not within the first camera view. We apply this mask to retain only the points that are not visible in the first camera's view, represented as:

$$\mathbf{P}_f = \begin{cases} \mathbf{P}, & \text{if } \mathbf{M} = 1 \\ 0, & \text{otherwise} \end{cases}, \tag{8}$$

and we get the filtered point cloud $\mathbf{P}_f$ of the moving scene. At the final stage of generation, we remove the overlapping regions of the other frames from the first frame's view,

obtaining a point cloud of the moving scene that is non-overlapping with the initial point cloud.

To keep the depth consistency between the moving scene and the static panorama scene, we use a depth scaler optimization strategy to maintain the depth consistency between the point cloud of the panorama scene and the moving scene. Specifically, we first convert the depth values into disparity, and then use a least squares-based optimization strategy to minimize the disparity difference between the first frame of the view-controlled video generated by DUSt3R and the corresponding disparity of the initial image obtained from the panorama depth estimation method. The optimization is represented as:

$$\min_{\alpha, \beta} \left\| \mathbf{M} \odot \left(\frac{\alpha}{d_p} + \beta - \frac{1}{d} \right) \right\|^2, \tag{9}$$

where α and β represent the scale and shift factors respectively. The mask $\mathbf{M}$ ensures that the depth alignment is conducted only in the overlapping regions, d represents the depth of the initial image of the moving scene and d_p represents the depth of the panorama in the corresponding region. The aligned depth is then given by:

$$\hat{d} = \left(\frac{\alpha}{\hat{d}_p} + \beta \right)^{-1}, \tag{10}$$

where $\hat{d}$ is the rectified depth of the initial image of the moving scene. Then we get the scale factor γ through $\gamma = \frac{\hat{d}}{d}$. Finally, we apply a 7×7 Gaussian kernel to smooth $\hat{d}$ at the mask edges, ensuring seamless transitions. For the depth of each subsequent frame d_i, we multiply with the same scale factor γ to get the rectified depth $\hat{d}_i$, ensuring that the point cloud of the moving scenes remain consistent in scale with the panorama point cloud.

Defined by users, the area that need to be expanded from the moving scene can be repeated, and we get the moving scenes $P_M^1, P_M^2, \cdots, P_M^i$. Finally, As both the static panorama scene and the moving scene are in the world coordinate system, we fuse the point cloud of the moving scenes with the panorama scene to get the final global-level point

cloud, represented as:

$$\Omega = \bigcup_{i=1}^{|P|} \varphi \left\{ P_0, P_M^1, \cdots, P_M^i \right\}, \qquad (11)$$

where φ represents the depth alignment function, P_0 represents the initial panorama point cloud, M_i represents the $i-th$ moving scene point cloud, and Ω represents the complete global-level point cloud.

4.4. Rendering with Refined 3D Gaussian Splatting

We finally use 3D point cloud as an accurate 3D representation. The generated global-level point cloud Ω serves as the initial Structure from Motion (SfM) points, which helps accelerate the convergence of the training. First, based on the base camera poses set P_B and the corresponding projected image set I_B, the set of moving camera poses P_M and the corresponding projected image set I_M, we conduct the initial densification process of the 3D Gaussian Splatting and the initial 3D Gaussians G_0. Although the initial densification process is able to fill in the missing hole of the 3D point cloud, there is misalignment in these regions. Therefore, we then use the supplementary camera set P_S, the supplementary image set I_S and the supplementary mask set M_S to refine this process. The supplementary images and poses provide additional supervision in the second stage of training until we get the final rectified 3D Gaussians G_1 with global-level consistent geometric consistency with high-quality details.

5. Experiments

In this section, we conduct experiments to evaluate our model's performance on both text-to-panorama generation and panoramic scene generation.

5.1. Experimental Setup

Evaluation Metrics For quantitative evaluation, to evaluate the non-reference quality of the rendered images in the scene, we render images using supplementary cameras, and employ traditional no-reference image quality assessment metrics: Natural Image Quality Evaluator (NIQE) [18] and Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [17]. We adopt QAlign [37], the state-of-the-art method in quality assessment benchmarks to evaluate the perceptual quality of image contents, which is divided into quality and aesthetic to evaluate the image quality and aesthetic quality respectively. We also use CLIP-T [23] score to evaluate the alignment between the text input and the generated perspective-of-view images.

Data For the text prompts used for text-to-3D reconstruction, we use GPT4 to generate random scene descriptions.

During evaluation, we use GPT4 to generate 40 prompts, including 20 indoor scenes and 20 outdoor scenes. The quantitative results are calculated through the average score of the scenes.

5.2. Main Results

For evaluating the quality of 3D scene generation, we compare our approach with state-of-the-art 3D scene generation methods: Text2Room [9], which employs an iterative mesh generation approach to represent the scene based on inpainting and monocular depth estimation, and LucidDreamer [6], which utilizes a warp-and-refine strategy to iteratively generate point clouds for novel views and subsequently employs 3D Gaussian Splatting (3D-GS) to obtain the Gaussians of the scene. Since LucidDreamer cannot directly generate 3D-GS from text prompts, we use Stable Diffusion v2.1 [24] to generate the initial conditioning image, ensuring consistency with our method. The comparison results are shown in Fig. 3 and Table 1.

The results indicate that Text2Room struggles to generate coherent scenes when style descriptions are included. Due to its render-refine-repeat scheme, Text2Room encounters alignment issues when there is significant variation between the generated images, which prevents the model from effectively distinguishing overlapping regions. This issue is particularly pronounced when the prompt contains numerous object descriptions. LucidDreamer, on the other hand, can only generate coherent scenes with limited camera movement. Due to the accumulation of geometry errors inherent to its warp-and-inpaint generation scheme, both Text2Room and LucidDreamer fail to maintain consistency between views, particularly during large camera movements. Consequently, these methods exhibit blurry boundaries and artifacts at the intersections between adjacent objects. Our method produces high-quality results with smooth boundaries and fewer artifacts in both indoor and outdoor scenes. Furthermore, our model achieves robust geometric consistency even under large camera movements, setting it apart from the compared approaches.

We present results for text-to-panorama generation compared with prior methods [1, 29] in Fig. 4. MultiDiffusion[1] directly generates panoramas using rectified diffusion, whereas MVDiffusion [29] first generates perspective-view images using diffusion models and then composes them into a panorama. The results demonstrate that with LLM guidance, our model effectively avoids generating duplicate objects and significantly enhances content diversity and generation quality.

We also visualize qualitative results in Fig. 5. The results show that the rendered images show accurate depth maps, which validates the accurate geometry of our rendered results.

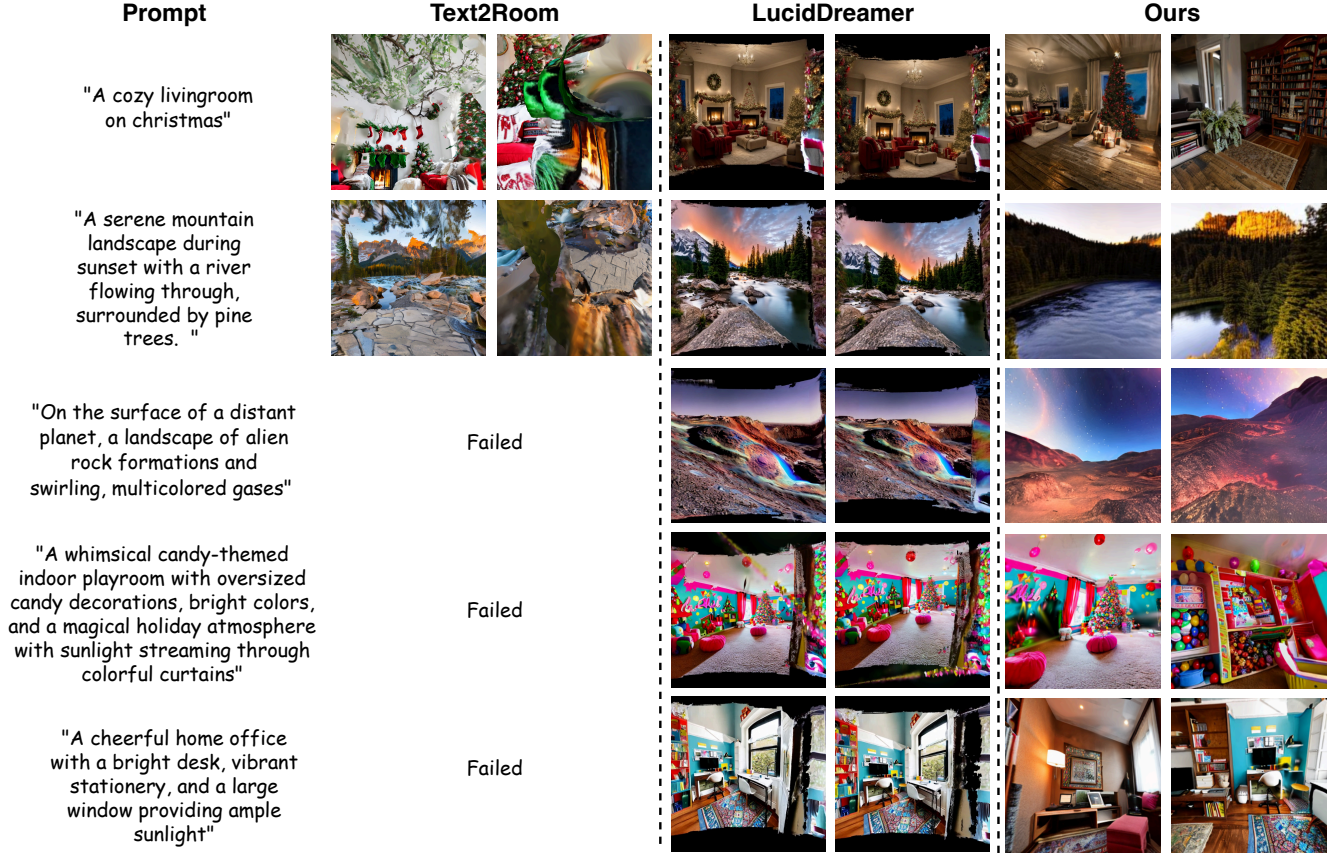| Prompt	Text2Room	LucidDreamer	Ours

Figure 3. Comparison of results. For Text2Room the images are projected from mesh and for LucidDreamer and our method, the images are projected from 3D Gaussians. Text2Room fail in generating specific stylized scenes. Compared with Text2Room and LucidDreamer, our method shows less artifacts and better geometry consistency.

Table 1. Comparison with other scene generation methods

	CLIP-T score ↑	Q-Align-Quality ↑	Q-Align-Aesthetic ↑	NIQE ↓	BRISQUE↓
Text2Room [9]	0.291	3.078	3.089	5.872	40.85
LucidDreamer [6]	0.296	3.051	3.035	6.132	45.78
Ours	**0.311**	**3.112**	**3.129**	**5.025**	**37.97**

Table 2. Ablation Study on rendered image quality.

	CLIP-T score ↑	Q-Align-Quality ↑	Q-Align-Aesthetic ↑	NIQE ↓	BRISQUE↓
w/o supplementary cameras	0.295	3.011	3.002	5.568	45.68
w/o depth alignment	0.302	3.035	3.028	5.156	40.39
Ours	**0.311**	**3.112**	**3.129**	**5.025**	**37.97**

5.3. Ablation Study

We conduct ablation studies to verify the supplementary camera set and the depth scale component. As shown in Table. 2, incorporating supplementary cameras and semantic-preserved generative warping enhances the refinement stage of 3D Gaussian Splatting, while also reducing artifacts in the rendered results. Removing the depth alignment module results in blending issues between scenes, causing pixel misalignment and increasing geometric deviations during 3D Gaussian generation. Since 3D-GS relies heavily on accurate point cloud initialization, incorporating depth alignment reduces the misalignment between the panorama scene and the moving scenes, ultimately improving the quality of the rendered images.

Figure 4. Comparison of Text-to-Panorama Generation. As panoramas generated by MultiDiffusion [1] and MVDiffusion [29] have both limited vertical FoV, for comparison, we only show our panorama before outpainting. Compared with previous methods, our method shows less duplicated objects and better generation quality.
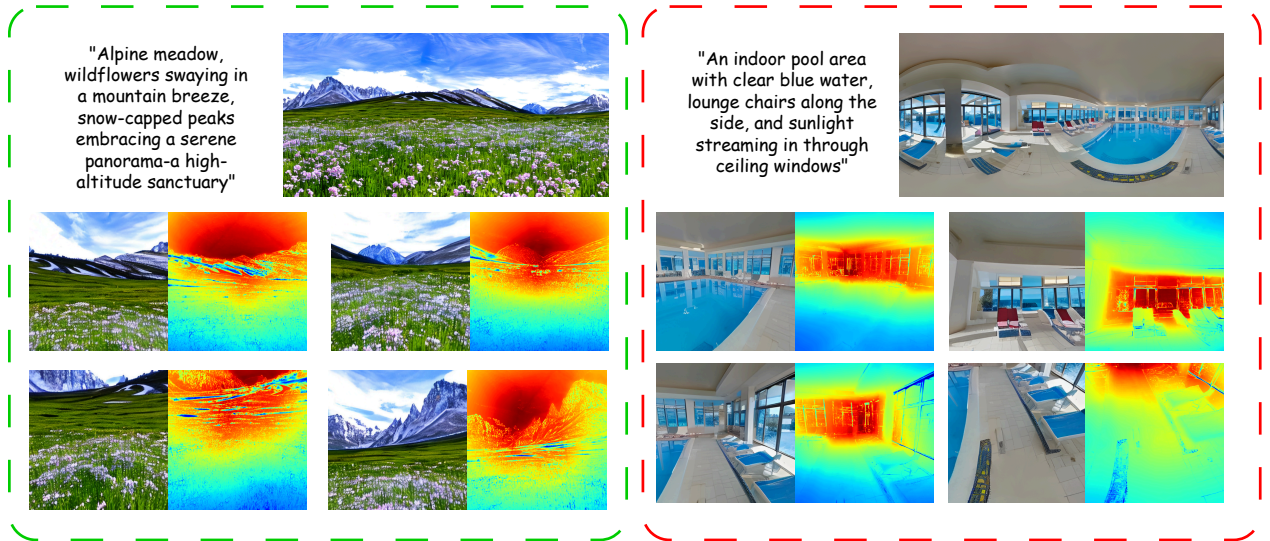

Figure 5. Our rendered rendered images with corresponding rendered depth.

We also compare the render quality of with other methods in Table 2. The results demonstrate that the exclusion of either the supplementary camera refinement or depth alignment leads to a significant degradation in rendering quality. These findings underscore the importance of both components in achieving high-quality scene reconstruction.

6. Conclusion

We proposed PanoDreamer, a text to 360-degree scene generation framework. The core insight of our method is to decompose scene generation into two phases: single-viewpoint scene generation and scene extension via moving camera simulation. The first phase uses an LLM to guide the synthesis of perspective images, which are then fused to form a panorama. During the second phase, the model is extended and improved using two different generation strategies. Our approach results in high-quality, geometry-consistent scenes, represented in the 3D-GS framework, and enables users to navigate freely along customized trajectories outside the initial, significantly broadening the range of potential applications. Our approach consistently outperforms strong baselines across a broad set of metrics. A key challenge that we plan to explore in future work is the accumulation of error as the scene scale becomes larger.

References

[1] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023. 2, 6, 8, 11, 12

[2] Zhipeng Cai, Matthias Mueller, Reiner Birkl, Diana Wofk, Shao-Yen Tseng, JunDa Cheng, Gabriela Ben-Melech Stan, Vasudev Lai, and Michael Paulitsch. L-magic: Language model assisted generation of images with coherence. In *CVPR*, 2024. 2, 3

[3] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation, 2023. 2

[4] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024. 2

[5] Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. Scene-dreamer: Unbounded 3d scene generation from 2d image collections. *IEEE transactions on pattern analysis and machine intelligence*, 45(12):15562–15576, 2023. 1

[6] Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv preprint arXiv:2311.13384*, 2023. 1, 2, 6, 7

[7] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *ICCV*, 2023. 2

[8] Mengyang Feng, Jinlin Liu, Miaomiao Cui, and Xuansong Xie. Diffusion360: Seamless 360 degree panoramic image generation based on diffusion models. *arXiv preprint arXiv:2311.13141*, 2023. 2

[9] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *ICCV*, 2023. 2, 6, 7

[10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2

[11] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 2

[12] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 2024. 1

[13] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *CVPR*, 2023. 2

[14] Gongye Liu, Menghan Xia, Yong Zhang, Haoxin Chen, Jinbo Xing, Xintao Wang, Yujiu Yang, and Ying Shan. Style-crafter: Enhancing stylized text-to-video generation with style adapter. *arXiv preprint arXiv:2312.00330*, 2023. 2

[15] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2

[16] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 2021. 2

[17] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21 (12):4695–4708, 2012. 6

[18] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. 6

[19] Muyao Niu, Xiaodong Cun, Xintao Wang, Yong Zhang, Ying Shan, and Yinqiang Zheng. Mofa-video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model. *arXiv preprint arXiv:2405.20222*, 2024. 2

[20] Elia Peruzzo, Vidit Goel, Dejia Xu, Xingqian Xu, Yifan Jiang, Zhangyang Wang, Humphrey Shi, and Nicu Sebe. Vase: Object-centric appearance and shape manipulation of real videos, 2024. 2

[21] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1

[22] Feng Qiao, Zhexiao Xiong, Eric Xing, and Nathan Jacobs. Genstereo: Towards open-world generation of stereo images and unsupervised matching. *arXiv preprint arXiv:2503.12720*, 2025. 1

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 2021. 6

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 4, 6

[25] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 1, 2

[26] Junyoung Seo, Kazumi Fukuda, Takashi Shibuya, Takuya Narihira, Naoki Murata, Shoukang Hu, Chieh-Hsin Lai, Seungryong Kim, and Yuki Mitsufuji. Genwarp: Single image to novel views with semantic-preserving generative warping. *arXiv preprint arXiv:2405.17251*, 2024. 4, 5

[27] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2

[28] Liangchen Song, Liangliang Cao, Hongyu Xu, Kai Kang, Feng Tang, Junsong Yuan, and Yang Zhao. Roomdreamer: Text-driven 3d indoor scene synthesis with coherent geometry and texture. *arXiv preprint arXiv:2305.11337*, 2023. 2

[29] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *arXiv*, 2023. 2, 6, 8, 11, 12

[30] Guangcong Wang, Peng Wang, Zhaoxi Chen, Wenping Wang, Chen Change Loy, and Ziwei Liu. Perf: Panoramic neural radiance field from a single panorama. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2024. 1, 2

[31] Hai Wang, Xiaoyu Xiang, Yuchen Fan, and Jing-Hao Xue. Customizing 360-degree panoramas through text-to-image diffusion models. In *WACV*, 2024. 2

[32] Ning-Hsu Wang and Yu-Lun Liu. Depth anywhere: Enhancing 360 monocular depth estimation via perspective distillation and unlabeled data augmentation. *arXiv preprint arXiv:2406.12849*, 2024. 4

[33] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 4

[34] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 2024. 2

[35] Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. Sinsr: diffusion-based image super-resolution in a single step. In *CVPR*, 2024. 4

[36] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *CVPR*, pages 7467–7477, 2020. 2

[37] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Chunyi Li, Liang Liao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtai Zhai, and Weisi Lin. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. Equal Contribution by Wu, Haoning and Zhang, Zicheng. Project Lead by Wu, Haoning. Corresponding Authors: Zhai, Guangtai and Lin, Weisi. 6

[38] Tianhao Wu, Chuanxia Zheng, and Tat-Jen Cham. Panodiffusion: 360-degree panorama outpainting via diffusion. In *The Twelfth International Conference on Learning Representations*, 2023. 4

[39] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*. Springer, 2025. 2

[40] Zhexiao Xiong, Xin Xing, Scott Workman, Subash Khanal, and Nathan Jacobs. Mixed-view panorama synthesis using geospatially guided diffusion. *arXiv preprint arXiv:2407.09672*, 2024. 2

[41] Zhexiao Xiong, Wei Xiong, Jing Shi, He Zhang, Yizhi Song, and Nathan Jacobs. Groundingbooth: Grounding text-to-image customization. *arXiv preprint arXiv:2409.08520*, 2024. 1, 2

[42] Shuai Yang, Jing Tan, Mengchen Zhang, Tong Wu, Yixuan Li, Gordon Wetzstein, Ziwei Liu, and Dahua Lin. Layerpano3d: Layered 3d panorama for hyper-immersive scene generation. *arXiv preprint arXiv:2408.13252*, 2024. 2

[43] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023. 2

[44] Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. Wonderjourney: Going from anywhere to everywhere. In *CVPR*, 2024. 1

[45] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 2, 4

[46] Ziyang Yuan, Mingdeng Cao, Xintao Wang, Zhongang Qi, Chun Yuan, and Ying Shan. Customnet: Zero-shot object customization with variable-viewpoints in text-to-image diffusion models. *arXiv preprint arXiv:2310.19784*, 2023. 1

[47] Cheng Zhang, Qianyi Wu, Camilo Cruz Gambardella, Xiaoshui Huang, Dinh Phung, Wanli Ouyang, and Jianfei Cai. Taming stable diffusion for text to 360◦ panorama image generation. In *CVPR*, 2024. 1

[48] Cheng Zhang, Qianyi Wu, Camilo Cruz Gambardella, Xiaoshui Huang, Dinh Phung, Wanli Ouyang, and Jianfei Cai. Taming stable diffusion for text to 360 {\deg} panorama image generation. *arXiv preprint arXiv:2404.07949*, 2024. 2

[49] Jingbo Zhang, Xiaoyu Li, Ziyu Wan, Can Wang, and Jing Liao. Text2nerf: Text-driven 3d scene generation with neural radiance fields. *arXiv preprint arXiv:2305.11588*, 2023. 2

[50] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 1, 2

[51] Haiyang Zhou, Xinhua Cheng, Wangbo Yu, Yonghong Tian, and Li Yuan. Holodreamer: Holistic 3d panoramic world generation from text descriptions. *arXiv preprint arXiv:2407.15187*, 2024. 2

[52] Shijie Zhou, Zhiwen Fan, Dejia Xu, Haoran Chang, Pradyumna Chari, Tejas Bharadwaj, Suya You, Zhangyang Wang, and Achuta Kadambi. Dreamscene360: Unconstrained text-to-3d scene generation with panoramic gaussian splatting. In *European Conference on Computer Vision*, 2025. 1, 2

Appendix

A. Experiment Details

During rendering, as we use pinehole cameras, reducing the camera Field-of-View(FoV) will reduce the distortions. During the projection, we set the camera FOV of the base camera and the supplementary cameras to be $60°$. We set the number of base camera to be 80, with each base camera corresponding to 4 supplementary cameras. The resolution of the projected images from base cameras and the supplementary cameras are set to be 512×512.

During the two-stage 3D gaussian splatting process, for the first 5000 iterations, we use the base set to initialize the 3D Gaussians, and after 5000 iterations, we add the refined supplementary set for the second-stage refinement of the 3D Gaussians. We report the performance of the rendered results on training 10,000 iterations.

B. Further Qualitative Results of Scene Generation

We show more qualitative results of our method on some scenes, shown in Fig. 6. Results show that our method not only generate high-quality rendered images, but also maintains accurate geometry and scene consistency.

C. Further Qualitative Results on Text-to-Panorama Generation

We present more results for text-to-panorama generation in comparison with prior methods [1, 29] in Fig. 7. Compared with previous methods, with LLM guidance, our method shows less duplicated objects and better generation quality.
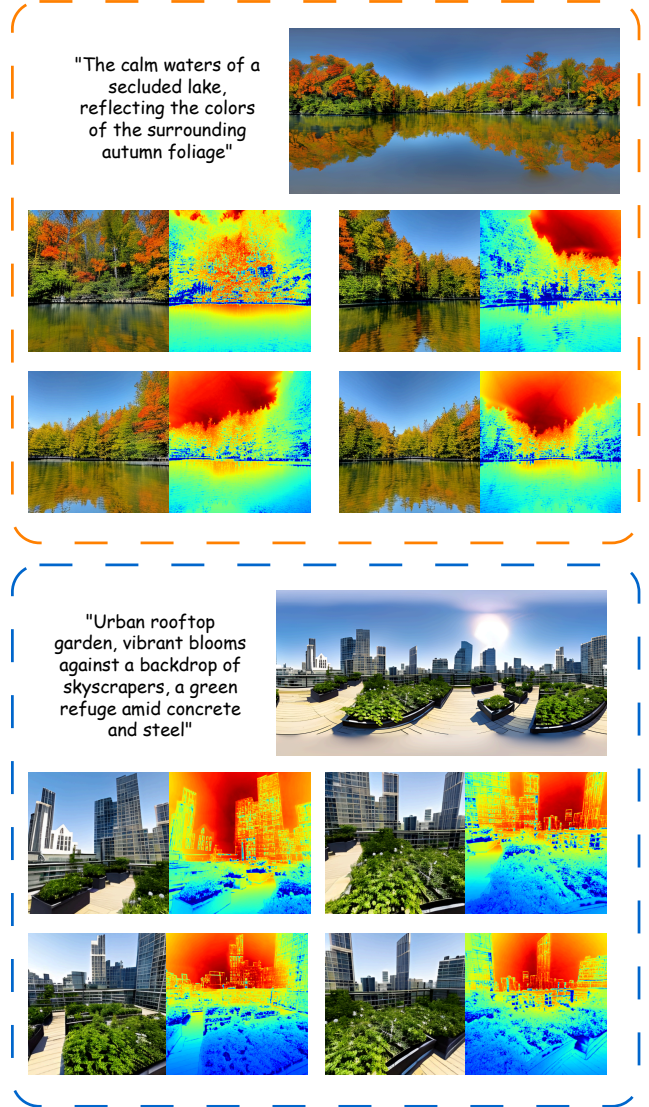

Figure 6. Additional results about scene generation. We show both the rendered images and rendered depth.

Figure 7. Additional comparison of text-to-panorama generation. As panoramas generated by MultiDiffusion [1] and MVDiffusion [29] have both limited vertical FoV, for a fair comparison, we only show our panorama before outpainting.