

# SceneDreamer360: Text-Driven 3D-Consistent Scene Generation with Panoramic Gaussian Splatting

Wenrui Li, Fucheng Cai, Yapeng Mi, Zhe Yang, Wangmeng Zuo, *Senior Member, IEEE*,  
Xingtao Wang, Xiaopeng Fan, *Senior Member, IEEE*

**Abstract**—Text-driven 3D scene generation has seen significant advancements recently. However, most existing methods generate single-view images using generative models and then stitch them together in 3D space. This independent generation for each view often results in spatial inconsistency and implausibility in the 3D scenes. To address this challenge, we proposed a novel text-driven 3D-consistent scene generation model: SceneDreamer360. Our proposed method leverages a text-driven panoramic image generation model as a prior for 3D scene generation and employs 3D Gaussian Splatting (3DGS) to ensure consistency across multi-view panoramic images. Specifically, SceneDreamer360 enhances the fine-tuned Panfusion generator with a three-stage panoramic enhancement, enabling the generation of high-resolution, detail-rich panoramic images. During the 3D scene construction, a novel point cloud fusion initialization method is used, producing higher quality and spatially consistent point clouds. Our extensive experiments demonstrate that compared to other methods, SceneDreamer360 with its panoramic image generation and 3DGS can produce higher quality, spatially consistent, and visually appealing 3D scenes from any text prompt. Our codes are available at <https://github.com/liwrui/SceneDreamer360>.

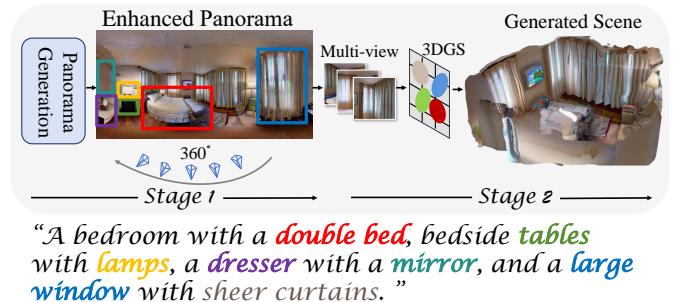
## I. INTRODUCTION

The scarcity of annotated 3D text-point cloud datasets poses substantial challenges to training 3D point cloud models directly from user queries, particularly for complex 3D scenes. This difficulty primarily stems from the high costs associated with acquiring and annotating 3D data, in addition to the precision required for modeling detailed 3D environments. To address these limitations, existing approaches frequently extend advanced 2D generative models by leveraging 2D object priors as a bridge to 3D neural radiance fields. For instance, CLIP-NeRF [1] uses the CLIP model [2] to control the shape and appearance of 3D objects based on textual prompts or images. Similarly, Dream Fields [3] integrates neural rendering with image and text representations to produce diverse 3D objects from natural language descriptions. Dreamfusion [4]

This work was supported in part by the National Key R&D Program of China (2021YFF0900500), the National Natural Science Foundation of China (NSFC) under grants 62441202, U22B2035, and the Fundamental Research Funds for the Central Universities under grants HIT.DZJJ.2024025. (Corresponding author: Xingtao Wang.)

Wenrui Li, Fucheng Cai, Yapeng Mi, Wangmeng Zuo, Xingtao Wang and Xiaopeng Fan are with the Department of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China. (e-mail: liwr@stu.hit.edu.cn;202111124@stu.hit.edu.cn;2021113246@stu.hit.edu.cn;wmzuo@hit.edu.cn;xtwang@hit.edu.cn;fxp@hit.edu.cn).

Zhe Yang is with the School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu, Sichuan 611731, China. (e-mail: 202221110110@std.uestc.edu.cn).



“A bedroom with a **double bed**, **bedside tables** with **lamps**, a **dresser** with a **mirror**, and a **large window** with sheer curtains.”

Fig. 1. We introduce SceneDreamer360, a text-based 3D scene generation framework designed to create realistic 3D scenes with high consistency across different viewpoints. SceneDreamer360 consists of two stages. In the first stage, the Panorama Generation module creates enhanced panoramic images to provide a complete and consistent spatial prior for the 3D scene. In the second stage, 3D Gaussian Splatting is used to reconstruct multi-view spatial images, resulting in a complete and spatially consistent point cloud.

further advances this approach by optimizing NeRF without the need for 3D training data, instead relying solely on a pre-trained 2D text-to-image diffusion model for text-to-3D synthesis.

Despite significant advancements in NeRF-based point cloud generation methods, achieving consistent, fine-grained point cloud details remains a challenge. For example, the Text2Room method [5] often produces point clouds that are discontinuous, lack detailed features, and require prolonged rendering times. Similarly, LucidDreamer [6] generates panoramic point clouds with numerous inconsistencies, resulting in incomplete and fragmented panoramas. PanFusion [7] also struggles with generating high-quality panoramic images from complex, extended text inputs, frequently yielding blurred areas and deformed objects. Moreover, PanFusion’s panoramic images are relatively low-resolution ( $512 \times 1024$ ), leading to a reduced spatial-to-pixel point ratio, which further contributes to blurred, suboptimal point cloud rendering. To address these challenges, we propose incorporating 3D Gaussian Splatting (3DGS) [8] for multi-scene generation, specifically aimed at producing more finely detailed and consistent complex scene point clouds. By employing a Gaussian distribution for scene modeling, 3DGS enables more precise representation of complex structures and improves reconstruction accuracy. Additionally, 3DGS’s optimized rendering algorithm significantly reduces rendering times while preserving high quality, resulting in faster and more efficient generation of

detailed point clouds.

In this paper, we introduce SceneDreamer360, a novel framework for text-driven 3D-consistent scene generation using panoramic Gaussian splatting (3DGS). Our method operates in two stages, as illustrated in Fig. 1: first, we enhance panoramic images to establish 3D consistency priors, and second, we apply 3DGS to reconstruct high-quality, text-aligned point clouds. This dual-stage approach ensures both accurate 3D structure and enhanced visual quality and realism in the generated scenes. During the panorama generation and enhancement stage, existing techniques frequently rely on iterative, progressive scene rendering, which can lead to consistency issues across successive renderings. To address this challenge, we enhance the PanFusion model [7] by integrating a multi-layer perceptron (MLP) [9] and a LoRA layer [10] into its final processing stage. These additions, combined with training on the Habitat Matterport Dataset [11], produce a model checkpoint fine-tuned to the specific demands of panoramic image generation. This enhancement step produces high-quality panoramas, forming a solid foundation for subsequent point cloud rendering. Further, we upscale and refine the generated panoramas using ControlNet (CN) [12] and Realesrgan [13] techniques, achieving resolutions up to 6K. This upscaling process preserves fine details and enhances image fidelity, which is crucial for realistic point cloud rendering. High-resolution panoramas contribute to more detailed and visually appealing scenes, which are important for achieving photorealism in the final output. To optimize the point cloud rendering process, we introduce a novel point cloud initialization method that improves 3D consistency and reduces rendering time. This step ensures better alignment of the resulting point clouds with the panoramic views, creating a more cohesive and immersive 3D experience. Additionally, the robust 3D representation capabilities of 3DGS enable the creation of complete, high-quality point cloud images that are consistent with the input text prompts. Our contributions can be summarised as follows:

- We propose SceneDreamer360, a domain-free high-quality 3D scene generation method, achieving better domain generalization in 3D scene generation by leveraging the panorama generation and 3D Guassian Splatting.
- Our method adapts panorama generation to provide a consistency prior for scenes, fine-tunes the model, and applies a three-stage enhancement to add more detail and spatial consistency to the final scene generation.
- To improve scene reconstruction, we employed 3D Gaussian Splatting and introduced a novel point cloud initialization and fusion algorithm, resulting in enhanced performance of the final generated point cloud scenes.

The remainder of this paper is organized as follows: Section II reviews the background and related work on 3D scene generation, 3D scene representation, and panorama generation. Section III describes the architecture and functionality of the proposed Scenedreamer360 model in detail. Section IV presents the experimental results and visualizations, highlighting the effectiveness of our approach. Finally, Section V summarizes the study's main findings and contributions.

## II. RELATED WORK

### A. 3D Scene Generation

The field of 3D scene generation has evolved significantly, drawing inspiration from various breakthroughs in image generation techniques. Early approaches leveraged Generative Adversarial Networks (GANs) [14], attempting to create multi-view consistent images [15], [16] or directly generate 3D representations like voxels [17], [18] and point clouds [19], [20]. However, these methods were hindered by GAN's inherent learning instability and the memory constraints of 3D representations, limiting the quality of generated scenes.

The advent of diffusion models [21], [22], [23] and their success in image generation [24], [25] sparked a new wave of research in 3D scene generation. Researchers began applying diffusion models to various 3D representations, including voxels [26], point clouds [27], and implicit neural networks [28]. While these approaches showed promise, they often focused on simple, object-centric examples due to their inherent nature.

To address more complex scenarios, some generative diffusion models employed meshes as proxies, diffusing in UV space. This approach enabled the creation of large portrait scenes through continuous mesh building [29] and the generation of indoor scenes [30], [31] and more realistic objects [32]. Currently, an increasing number of studies [6], [33], [34], [35] have incorporated 3DGS [8] as a 3D representation to complement diffusion models in generating more complex 3D scenes. For instance, LucidDreamer [6] pioneered the integration of 3D Gaussian Splatting into scene generation, demonstrating the efficacy of 3DGS. RealmDreamer [36] optimizes a 3D Gaussian Splatting representation to align with complex text prompts. However, due to diffusion models generating one viewpoint at a time before merging and elevating to 3D space, the resulting 3D scenes often lack realism and consistency. To address this issue, some recent methods such as DreamScene360 [33] and FastScene [34] have introduced panoramic images to generate more coherent 3D scenes. Nevertheless, these approaches still produce scenes lacking in detail and quality, primarily due to limitations in panoramic image generation models and view fusion algorithms.

In contrast, to address the issues of quality and spatial consistency, our method leverages the promising 3D Gaussian Splatting technique and implements a three-stage enhancement process for the generated panoramic images. Concurrent work HoloDreamer [37] has explored similar concepts. However, our approach is distinguished by an innovative design in the panoramic image enhancement stage and a novel point cloud fusion algorithm. Our approach improves the overall quality and spatial coherence of the generated 3D scenes, advancing the state-of-the-art in text-driven 3D scene generation.

### B. 3D Scene Representation

The field of 3D scene representation [38], [39] has evolved significantly, encompassing a variety of approaches with distinct trade-offs and applications. Traditional explicit methods such as point clouds, meshes, and voxels have long been fundamental to 3D modeling, offering intuitive control and fast rendering through rasterization. However, these methods often

require a large number of elements to achieve high resolution, which can be computationally expensive. To address this limitation, more complex primitives are developed, including cuboids [40], [41], [42] and polynomial surfaces [43], offering more efficient expression of geometry but still struggling with realistic color representation.

Recent advancements have led to the emergence of implicit neural representations, such as Signed Distance Functions (SDF) [44] and Neural Radiance Fields (NeRF) [45]. These methods, particularly NeRF, have demonstrated the ability to represent complex 3D shapes and textures with rich details. However, they face challenges in terms of handling and optimization efficiency. To improve upon these limitations, subsequent research has focused on combining volume rendering with explicit representations, utilizing structures like sparse voxels [46], [47], [48], [49], featured point clouds [50], Multi-Level Hierarchies [51], [52], tensors [53], infinitesimal networks [54], [55], triplanes [56], and polygons [57].

Among these advancements, 3D Gaussian Splatting(3DGS) [8] has emerged as a particularly promising approach. It represents complete and unbounded 3D scenes by effectively 'splatting' Gaussians, utilizing spherical harmonics and opacity for strong representational capabilities. 3DGS balances quality and efficiency by supporting alpha-blending and differentiable rasterization, enabling fast 3D scene optimization. It refines point clouds details through a split-and-clone mechanism, making it ideal for complex scene generation where scene bounds are uncertain.

### C. Panorama Generation

Panoramas offer a comprehensive, unobstructed view that captures extensive scene areas. Recent advancements in panorama generation have significantly improved the quality and consistency of 360-degree images. While early attempts like PanoGen [58] and MultiDiffusion [59] used pre-trained diffusion models to create long images from text prompts, these methods often resulted in inconsistencies and lacked true panoramic properties. To address these limitations, researchers have developed more sophisticated approaches. MVDiffusion [60] introduced a correspondence-aware attention mechanism to generate consistent multi-view images simultaneously. This method ensures that the different views maintain coherence and alignment, which is crucial for creating seamless panoramic images. However, MVDiffusion still faced challenges in achieving perfect continuity at the image edges. StitchDiffusion [61] improved upon it by using LoRA [10] fine-tuning to generate complete 360-degree panoramas with enhanced continuity between image edges. This approach leverages the Low-Rank Adaptation (LoRA) technique to fine-tune the model, ensuring that the transitions between different parts of the panorama are smooth and natural. PanFusion [7] proposed a novel dual-branch diffusion model to generate a 360-degree image from a text prompt. This model utilizes two separate branches to handle different aspects of the image generation process, allowing for more detailed and context-aware outputs. The dual-branch architecture helps in capturing both the global structure and local details of the scene, resulting in high-quality panoramic images.

In this paper, we incorporated MLP and LoRA into the PanFusion model and performed fine-tuning to enhance the consistency of the generated panoramic images. By integrating a Multi-Layer Perceptron (MLP) with LoRA in the PanFusion framework, we aim to further improve the coherence and quality of the panoramas. The MLP component helps in refining the details and ensuring that the generated images are not only consistent but also rich in detail. Furthermore, recent studies have explored the integration of machine learning techniques with traditional computer vision methods to further enhance panorama generation. For instance, hybrid models that combine deep learning with geometric alignment techniques have shown promise in reducing artifacts and improving the overall visual quality of the panoramas. These advancements highlight the ongoing efforts to push the boundaries of what is possible in panorama generation, making it an exciting and rapidly evolving field. Overall, the integration of advanced attention mechanisms, fine-tuning techniques, and hybrid models has significantly advanced the state of panorama generation, enabling the creation of high-quality, seamless 360-degree images from text prompts.

## III. METHOD

This method introduces a novel pipeline for generating high-resolution, 360-degree panoramic point cloud scenes from arbitrary spatial text inputs. The pipeline consists of two main stages: Panoramic Image Generation and Processing and Point Cloud Construction with Gaussian Splatting. As illustrated in Fig. 2, the SceneDreamer360 framework effectively combines these stages to produce visually coherent and spatially consistent 3D scenes, capturing fine details and maintaining high image fidelity.

In the first stage, an adapted PanFusion [7] model generates initial panoramic images, which are subsequently refined through a three-step super-resolution process, enhancing both image details and resolution. Next, we use the high-resolution panorama as input and apply the Equirectangular to Perspective (e2p) algorithm to obtain multiple scene images from different perspectives. In the second stage, point clouds are constructed using a 3D Gaussian Splatting [8] technique. This process involves initializing the point cloud, applying masking to remove duplicates for a more complete map, and training a 3D Gaussian splatting model along a new rendering trajectory to produce a detailed 3D spatial scene point cloud.

### A. Panoramic Image Generation and Processing

The goal of Panoramic Image Generation and Optimization is to convert a long-text description of a complex scene into high-quality 360-degree panoramas with high resolution. The specific steps are as follows:

*1) Complex Long-Text Generation:* We used the GPT-4 model to generate a series of similar complex scene descriptions to increase the diversity and richness of the training data. The user first inputs an initial long text description of a complex scene. To generate more texts with similar semantics and content, we utilize GPT-4 to expand and transform it. These generated texts cover different details and perspectives,

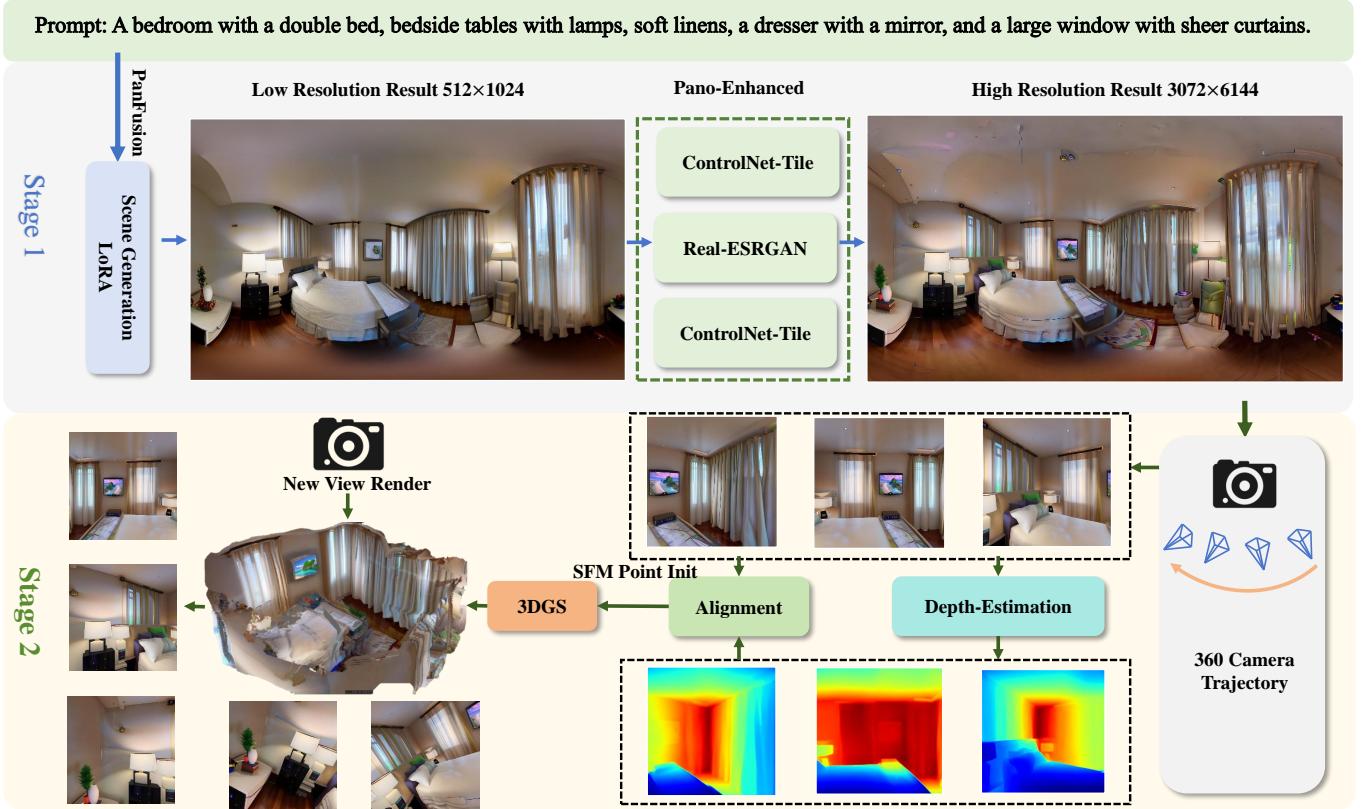


Fig. 2. The architecture of the SceneDreamer360. SceneDreamer360 generates an initial panorama from any open-world textual description using the fine-tuned PanFusion model. This panorama is then enhanced to produce a high-resolution  $3072 \times 6144$  image. In the second stage, multi-view algorithms generate multi-view images, and a monocular depth estimation model provides depth maps for initial point cloud fusion. Finally, 3D Gaussian Splatting is used to reconstruct and render the point cloud, resulting in a complete and consistent 3D scene.

making the subsequent panorama generation more diverse and comprehensive.

2) *Finetune PanFusion Model*: Given the input long text representation  $T$  and its embedding  $e_T$ , the representation after the MLP layer is:

$$\mathbf{h} = \sigma(\mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1 \cdot \mathbf{e}_T + \mathbf{b}_1) + \mathbf{b}_2), \quad (1)$$

where  $W_1$  and  $W_2$  are the weight matrices of the MLP layer,  $b_1$  and  $b_2$  are the bias terms, and  $\sigma$  is the activation function.

### B. Panoramic Image Generation

Our method utilized the PanFusion model [7] to generate panoramas. However, because the PanFusion model was originally trained on simple short text prompts, it required appropriate adjustments and fine-tuning to handle complex long texts. Therefore, we added a Multi-Layer Perceptron (MLP) [9] layer and a Low-Rank Adaptation (LoRA) [10] layer to the original PanFusion model, fine-tuning it using the reannotated Habitat Matterport Dataset [11]. The MLP layer captures higher-order features in complex long texts, while the LoRA layer enhances the model's representation ability through low-rank matrix decomposition. For model fine-tuning, we first reannotated the dataset using the generated complex long texts, pairing them with corresponding panoramas to form new training samples. Next, we froze the pre-trained parts of the PanFusion model and trained only the newly added MLP and LoRA layers. This

ensures that the model effectively utilizes information from complex long texts to generate high-quality panoramas. These panoramas encapsulate the details and scenes described in the input text, serving as foundational data for subsequent steps.

Given the input long text representation  $T$  and its embedding  $e_T$ , the representation after the MLP layer is:

$$\mathbf{h} = \sigma(\mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1 \cdot \mathbf{e}_T + \mathbf{b}_1) + \mathbf{b}_2), \quad (2)$$

where  $W_1$  and  $W_2$  are the weight matrices of the MLP layer,  $b_1$  and  $b_2$  are the bias terms, and  $\sigma$  is the activation function.

The representation of the LoRA layer is:

$$\mathbf{h}' = \mathbf{h} + \mathbf{U} \cdot \text{softmax}(\mathbf{V} \cdot \mathbf{h}), \quad (3)$$

where  $U$  and  $V$  are the low-rank matrices of the LoRA layer.

1) *Dataset Reannotation*: We reannotated the dataset using the generated complex long texts. Specifically, we paired the complex long texts with corresponding panoramas to form new training samples.

2) *Model Fine-tuning*: During fine-tuning, we froze the pre-trained parts of the PanFusion model and only trained the newly added MLP and LoRA layers. This ensures that the model can effectively utilize information from the complex long texts to generate high-quality panoramas.

The adjusted and fine-tuned PanFusion model is applied to the generated long text collection, producing 360-degree

panoramas. These panoramas contain details and scenes described in the input text and serve as the foundational data for subsequent steps.

### C. Panoramic Image Optimization

As the pre-trained PanFusion model can only generate panoramas with a resolution of  $512 \times 1024$ , directly using these panoramas for multi-view extraction and point cloud reconstruction would result in sparse point clouds with blurry 3D Gaussian rendering [8]. To meet the requirements of point cloud reconstruction, it is necessary to further enhance the resolution and quality of the panoramas. The specific steps are as follows: We adopted a three-step super-resolution reconstruction method involving ControlNet-Tile [12], RealESRGAN [13], and a second application of ControlNet-Tile to maximize the enhancement of image details and resolution.

1) *ControlNet-Tile Super-Resolution Reconstruction*: The ControlNet-Tile model divides the panorama into distinct tiles through semantic recognition. Each tile is analyzed by the model to determine which details to add and which parts to repair. Based on tiling, ControlNet-Tile enhances the details of each tile. The model can identify various areas in the image and enhance the details of each area, adding additional detail where necessary. The detail enhancement process can be expressed as:

$$\mathbf{I}_{enhanced} = f_{tile}(\mathbf{I}_{ori}, t), \quad (4)$$

where  $f_{tile}$  represents the ControlNet-Tile model,  $\mathbf{I}_{ori}$  represents the original panorama, and  $t$  represents the input text.

2) *RealESRGAN Super-Resolution Reconstruction*: Based on ControlNet-Tile, we further employed RealESRGAN for higher precision super-resolution reconstruction. RealESRGAN is an advanced super-resolution method based on generative adversarial networks, capable of enhancing resolution while preserving natural details and visual consistency. The super-resolution reconstruction process can be expressed as:

$$\mathbf{I}_{super} = g_{ESRGAN}(\mathbf{I}_{enhanced}), \quad (5)$$

where  $g_{ESRGAN}$  represents the RealESRGAN model.

3) *ControlNet-Tile Re-Optimization*: After RealESRGAN processing, we applied ControlNet-Tile once more for additional refinement and optimization. The purpose of this second application of ControlNet-Tile is to further enhance the details and quality of the image at the increased resolution. The final optimized image  $\mathbf{I}_{final}$  is represented as:

$$\mathbf{I}_{final} = f_{tile}(\mathbf{I}_{super}, t). \quad (6)$$

By following these steps, we increased the panorama's resolution to  $3072 \times 6144$ . This high-resolution panorama preserves the original image's overall structure while significantly enhancing details and clarity, thus providing high-quality input data for subsequent point cloud reconstruction.

### D. Multi-View Image Acquisition

After obtaining the high-resolution panorama, we need to extract images from specific viewpoints for point cloud reconstruction. We select a suitable set of camera poses  $P_i \in \mathbb{R}^{4 \times 4}$

based on the requirements of point cloud reconstruction. The selection of these camera poses is based on geometric and visual coverage principles, ensuring that the chosen perspectives comprehensively cover important details and areas of the panorama. Specifically, we use an optimization-based viewpoint selection algorithm that considers the distribution of viewpoints, parallax effects, and coverage of reconstruction areas. The camera pose  $P_i$  is represented as:

$$\mathbf{P}_i = \begin{bmatrix} \mathbf{R}_i & \mathbf{T}_i \\ \mathbf{0} & 1 \end{bmatrix}. \quad (7)$$

The rotation matrix  $R_i$  is expressed as:

$$\mathbf{R}_i = \begin{bmatrix} \cos(\theta) & 0 & \sin(\theta) \\ 0 & 1 & 0 \\ -\sin(\theta) & 0 & \cos(\theta) \end{bmatrix}. \quad (8)$$

The translation matrix  $T_i$  is expressed as:

$$\mathbf{T}_i = \begin{bmatrix} T_x \\ T_y \\ T_z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}. \quad (9)$$

By using the Equirectangular to Perspective (e2p) algorithm, we transformed the panorama to generate perspective images from specific viewpoints. The e2p algorithm converts spherical panoramas into multiple planar images from different viewpoints, thereby creating parallax between views. The perspective transformation formula is:

$$\mathbf{I}_{perspective}(\mathbf{u}, \mathbf{v}) = \mathbf{E}(\mathbf{P}(f(\mathbf{u}, \mathbf{v}), \mathbf{R}_i, \mathbf{T}_i), \mathbf{S}), \quad (10)$$

where  $f(\mathbf{u}, \mathbf{v})$  represents the mapping function from planar coordinates  $(\mathbf{u}, \mathbf{v})$  to spherical coordinates,  $\mathbf{S}$  represents the specified output size,  $\mathbf{E}$  represents the view size transformation function.

Based on the camera poses, the e2p algorithm generates perspective images from the panorama. These perspective images retain the high resolution and details of the panorama while introducing parallax between views, providing rich perspective information for subsequent point cloud reconstruction.

By employing these methods, we can effectively generate high-quality, 360-degree panoramas directly from text prompts. This approach provides high-resolution and detailed input data, forming an optimal basis for the subsequent point cloud reconstruction. Using precise perspective transformations, we produce a series of viewpoint images that are well-suited to the requirements of point cloud reconstruction. These images not only capture various angles but also retain the visual fidelity and spatial coherence needed for constructing accurate 3D scenes. Consequently, they serve as a solid foundation for the point cloud reconstruction process, enhancing the realism and consistency of the final generated scene.

### E. Point Cloud Construction with Gaussian Splatting

After obtaining high-quality multi-view panoramic images, we apply the 3D Gaussian Splatting method for point cloud generation. We directly use 3D Gaussians to represent point clouds and have designed a novel point cloud alignment method. Additionally, we improved the point cloud initialization by applying masking to remove duplicates, resulting in a

more refined and complete initial point cloud map. The point cloud reconstruction process is divided into two phases: Point Cloud Initialization and Rendering.

1) *Point Cloud Initialization*: The initial point cloud must be generated from multi-view images to enable reconstruction using 3D Gaussian Splatting [8]. To generate a 360-degree spatial point cloud from multi-view images, a rendering trajectory centered around the observation viewpoint is required. This trajectory should cover the bottom, middle, and top of the spatial area. Using the LucidDreamer [6] framework, we incorporate additional camera viewpoints to create a more accurate spatial point cloud. For the set of multi-view images  $\mathbf{I}_1, \mathbf{I}_2, \dots, \mathbf{I}_n$  derived from the panoramic image, each image  $\mathbf{I}_i$  has a corresponding position  $\mathbf{P}_i$ . Using the monocular depth estimation model ZoeDepth [62], we measure the depth map  $\mathbf{D}_i$  for each  $\mathbf{I}_i$ , forming the depth map set  $\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_n$ . We initialize the point cloud using a cyclic generation approach. The camera's intrinsic matrix and the extrinsic matrix of  $\mathbf{I}_0$  are denoted as  $\mathbf{K}$  and  $\mathbf{P}_0$ , respectively. The initial point cloud for a single viewpoint is derived using the following formula:

$$\rho_0 = \tau_{2 \rightarrow 3}(\mathbf{I}_0, \mathbf{D}_0, \mathbf{K}, \mathbf{P}_0), \quad (11)$$

where  $\tau_{2 \rightarrow 3}(\cdot)$  is a function that elevates a 2D image to a 3D point cloud. This function generates an initial point cloud  $\rho_i$ , corresponding to a viewpoint  $P_i$ .

For point cloud fusion from different perspectives, we employed a masking method to remove duplicates. Specifically, we defined a mask  $\mathbf{M}_i$  for each generated  $\rho_i$ , where the mask projects the already generated  $\rho_{i-1}$  onto the 2D image space through the corresponding view  $\mathbf{P}_i$ . The projected part of the mask has values greater than 0, while the rest remains black with a value of 0. At the final stage of generation, we remove the duplicated parts that have already been projected, effectively merging the point clouds from different perspectives and reducing duplicate points. Meanwhile, we estimate the optimal depth scaling factor  $d_i$  to minimize the distance between the 3D points of the new image and the corresponding points in the original point cloud  $\mathbf{P}_{i-1}$ . Then, by multiplying the factor  $d_i$  with the measured depth map  $\mathbf{D}_i$ , we calculate the corrected depth map  $d_i\mathbf{D}_i$ , which is subsequently used for the generation of  $\rho_i$ . The complete process is illustrated in the following formula:

$$d_i = \operatorname{argmin}_d \left( \sum \|\tau_{2 \rightarrow 3}([\mathbf{I}_i, d\mathbf{D}_i], \mathbf{K}, \mathbf{P}_i) - \rho_{i-1}\|_1 \right), \quad (12)$$

$$\rho_i = \tau_{2 \rightarrow 3}(\mathbf{I}_i - \mathbf{M}_i, d_i\mathbf{D}_i, \mathbf{K}, \mathbf{P}_i), i \neq 0. \quad (13)$$

Finally, through continuous iteration of the formula, we fuse the point clouds to form the final point cloud  $\Omega$ .

$$\Omega = \bigcup_{i=1}^{|P|} \varphi \{\rho_i, \rho_{i-1}\}. \quad (14)$$

The function  $\varphi$  represents the depth alignment function, and all  $\rho$  are integrated to obtain the complete initial point cloud  $\Omega$ . **The overall algorithm is shown in Algorithm 1.**

2) *Rendering with Gaussian Splatting*: 3D Gaussian Splatting is a technique that efficiently models 3D scenes by leveraging a fast and accurate point-based representation. Following the generation of the initial point cloud, we use this point

---

**Algorithm 1** Point cloud initialization algorithm

---

**Input:** Panoramic multiview RGBD images  $\{\mathbf{I}_i, \mathbf{D}_i\}_{i=0}^N$

**Input:** Camera intrinsic  $\mathbf{K}$ , extrinsics  $\{\mathbf{P}_i\}_{i=0}^N$

**Output:** Complete point cloud  $\Omega$

```

1:  $\rho_0 = \tau_{2 \rightarrow 3}(\mathbf{I}_0, \mathbf{D}_0, \mathbf{K}, \mathbf{P}_0)$ 
2: for  $i = 1$  to  $N$  do
3:    $d_i = 1$ 
4:    $\mathbf{M}_i = \psi_{3 \rightarrow 2}(\rho_{i-1}, \mathbf{K}, \mathbf{P}_i)$ 
5:   while not converged do
6:      $\tilde{\rho}_i = \tau_{2 \rightarrow 3}(\mathbf{I}_i, d_i\mathbf{D}_i, \mathbf{K}, \mathbf{P}_i)$ 
7:      $\mathcal{L}_d \leftarrow \frac{1}{\|\mathbf{M}_i\|} \sum_{\mathbf{M}_i=1} \|\tilde{\rho}_i - \rho_{i-1}\|_1$ 
8:     Calculate  $\nabla_d \mathcal{L}_d$ 
9:      $d_i \leftarrow d_i - \alpha \nabla_d \mathcal{L}_d$ 
10:    end while
11:    $\rho_i = \tau_{2 \rightarrow 3}(\mathbf{I}_i - \mathbf{M}_i, d_i\mathbf{D}_i, \mathbf{K}, \mathbf{P}_i)$ 
12:    $\Omega = \rho_{i-1} \bigcup \rho_i$ 
13: end for
14: return  $\Omega$ 

```

---

cloud to train a 3D Gaussian Splatting model. The generated point cloud, denoted as  $\Omega$ , serves as the initial Structure from Motion (SfM) points, which accelerates network convergence and encourages the model to focus on producing detailed and realistic representations of the scene. To train the model, we utilize multi-view image slices as training images and incorporate the image view augmentation technique from LucidDreamer to enrich the data. Since the input multi-view images may not fully cover the entire 3D space, we project the  $\Omega$  point cloud into 2D from various camera viewpoints, effectively increasing the number of supervisory images and expanding spatial coverage. This augmentation process helps to provide more comprehensive training data, ensuring better spatial consistency and detail across the 3D scene.

$$\mathbf{I}_i, \mathbf{M}_i = \psi_{3 \rightarrow 2}(\Omega, \mathbf{K}, \mathbf{P}_i), \quad (15)$$

where  $\psi_{3 \rightarrow 2}(\cdot)$  is the function that projects the 3D point cloud onto the 2D plane. Through this method, we achieve a more consistent and complete scene point cloud.

## IV. EXPERIMENTS

In this section, we conducted a thorough and comprehensive evaluation of our method to validate its superiority. We performed both qualitative and quantitative evaluations that our method can generate clearer and higher quality 3D scenes.

### A. Implementation Details

In the first stage, we generate high-quality panoramic images using a modified PanFusion model, which has been fine-tuned on the Habitat Matterport Dataset. To enhance its capacity for producing realistic 3D panoramic images, we incorporate MLP and LoRA layers into the model. This fine-tuning enables the generation of more visually coherent and lifelike panoramas. We further apply a three-step super-resolution process, using ControlNet-Tile and RealESRGAN, which enhances the resolution of these panoramas to  $3072 \times$

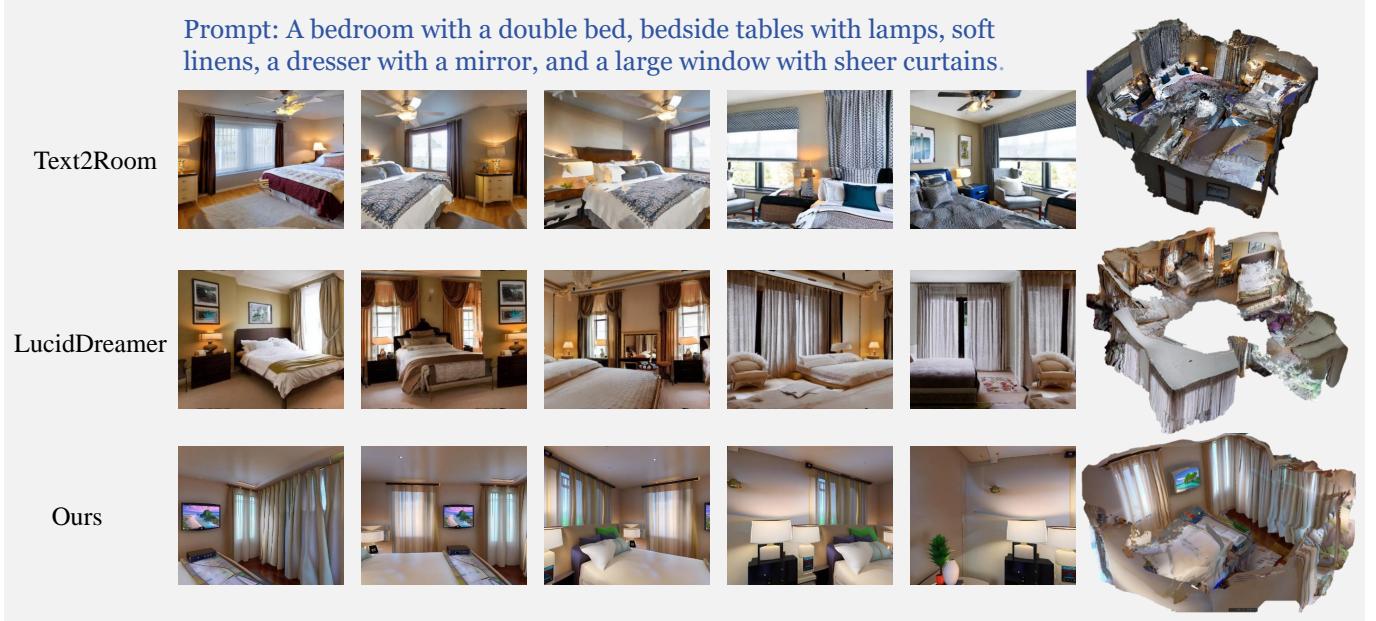


Fig. 3. The visualization comparison of the proposed SceneDreamer360 between current methods. The images on the left display the new viewpoint renderings, demonstrating how each method handles various perspectives. In contrast, the images on the right present the spatial views of the generated 3D scenes, allowing a direct comparison of the spatial consistency and overall structure of the reconstructions.

6144, preserving fine details and enriching the image quality for subsequent stages.

In the second stage, we adapt and modify the point cloud reconstruction component from LucidDreamer. This adaptation allows us to effectively convert the high-resolution panoramic images into detailed 3D point clouds, ensuring consistency with the input scene and text prompt. This two-stage process of panorama generation and point cloud reconstruction is designed to maximize visual fidelity and structural accuracy in the final 3D scenes.

Our method is implemented using PyTorch. We obtain camera viewpoints from given camera poses and use the e2p algorithm to extract corresponding multi-view images from the panorama. The specific implementation method is as follows:

```

1 for i in range(len(render_poses)):
2     print(i)
3     R=render_poses[i, :3, :3].T
4     T==render_poses[i, :3, 3:4]
5     output_size=(img_size,img_size)
6     pers_img=get_pers_img(
7         pano_img_path,R,T,
8         output_size=output_size)
9     pers_img_pil=Image.fromarray(
10        cv2.cvtColor(pers_img,
11                      cv2.COLOR_BGR2RGB))
12     img_list.append(pers_img_pil)

```

For evaluation, we generate a diverse set of prompts using GPT-4, covering a broad spectrum of scene types to comprehensively test our model's versatility. We then assess the quality of our outputs using multiple metrics, including CLIP-Score, PSNR, and others, to evaluate both visual similarity and fidelity to the input text prompts. All experiments are conducted on an NVIDIA A100-80G GPU, ensuring sufficient computational power for handling high-resolution image pro-

cessing and point cloud generation.

### B. Baselines

We compare our SceneDreamer360 with the following state-of-the-art baselines:

- TextToRoom [5] employs a mesh-based 3D representation, directly extracting meshes from inpainted RGBD images to represent watertight indoor environments.
- LucidDreamer [6] projects outpainted RGBD images onto a point cloud and then supervises the optimization of 3DGS by projecting multiple images from this point cloud. However, since LucidDreamer cannot directly generate 3D scenes from prompts, we utilize a diffusion model to generate condition images, allowing LucidDreamer to create scenes from text input.

**Evaluation Metrics** As there are currently no established evaluation metrics specifically for 3D scene point clouds, we assess our method using metrics adapted from prior work in related areas. We calculate the CLIP-Score [63] to measure text-image consistency, allowing us to evaluate how well the generated images align with the input text prompts. To assess the visual quality of single viewpoint renderings, we use CLIP-IQA [64], which provides an image quality evaluation aligned with human perception. Additionally, we employ PSNR, SSIM, and LPIPS to evaluate the fidelity and perceptual quality of the rendered images. These metrics collectively enable a well-rounded assessment, capturing both the accuracy of text-image correspondence and the visual quality of the generated 3D scenes.

### C. Qualitative Result

The point cloud images generated by our method using various prompts are displayed in Fig. 4. These results demonstrate

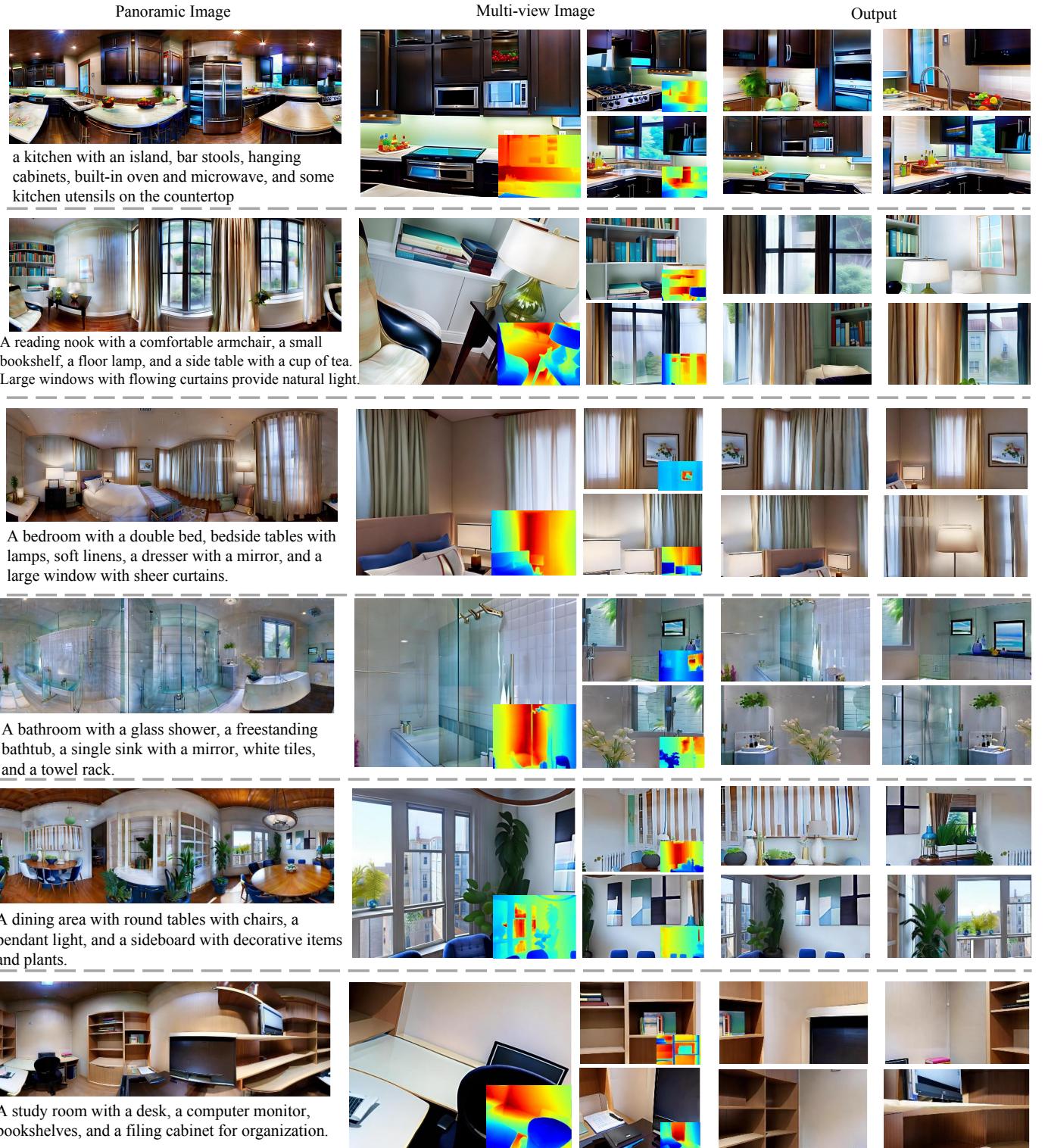


Fig. 4. The Panoramic Image on the left represents the high-resolution enhanced panorama generated from a specific text input, showcasing the detailed and cohesive scene captured in 360 degrees. The Multi-view Image in the center displays various perspective images derived from this panorama, illustrating the scene's adaptability across different viewpoints. Finally, the Output on the right presents images rendered from the generated point cloud at new viewpoints, highlighting SceneDreamer360's ability to maintain high quality and spatial consistency in the 3D reconstructions. As shown in the figure, SceneDreamer360 effectively produces detailed, visually coherent, and spatially consistent 3D scenes.

TABLE I  
COMPARISON OF IMAGE AND RENDER QUALITY ACROSS DIFFERENT METHODS.

Method	Image Quality				Render Quality		
	CLIP-Score↑	Sharp↑	Colorful↑	Resolution↑	PSNR↑	SSIM↑	LPIPS↓
Text2Room [5]	-	-	-	-	19.29	0.728	0.206
LucidDreamer [6]	0.2110	0.9437	0.3230	0.4757	29.56	0.922	0.039
<b>SceneDreamer360</b>	<b>0.2534</b>	<b>0.9877</b>	<b>0.3602</b>	<b>0.7082</b>	<b>35.74</b>	<b>0.963</b>	<b>0.035</b>

TABLE II  
QUANTITATIVE COMPARISON OF DIFFERENT SETTINGS.

Method	PSNR↑	SSIM↑	LPIPS↓
<b>Standard</b>	<b>34.96</b>	<b>0.962</b>	0.037
W/o enhance	30.36	0.947	<b>0.011</b>
W/o alignment	10.30	0.365	0.732
W/o enhance & alignment	9.70	0.281	0.803

that our method effectively creates 3D spatial scenes that are closely aligned with the input text prompts. Additionally, we show that the generated images can be rendered from arbitrary viewpoint trajectories, highlighting the generalizability and flexibility of our approach. To further evaluate our method, we visually compared the output with LucidDreamer and Text2Room using identical spatial text prompts, as illustrated in Fig. 3. Our method surpasses these alternatives by producing more consistent, complete, and detailed 3D spatial scene point clouds. Specifically, the integration of 3D Gaussian Splatting (3DGS) enables our model to capture finer details and maintain continuous object surfaces, resulting in sharper, more defined images compared to Text2Room, which struggles with clarity and fine details.

Our Stage 1 panoramic optimization also contributes significantly by capturing extensive spatial details, while our tailored rendering trajectory produces comprehensive, cohesive scenes. In contrast, LucidDreamer frequently generates point clouds that are incomplete, with visible voids that compromise the spatial coherence of the scene. Text2Room, on the other hand, often fails to achieve clarity, particularly in regions such as floors and other large surfaces, leading to a lack of detail. In comparison, SceneDreamer360 consistently delivers aesthetically pleasing and well-detailed indoor scenes that accurately reflect the text prompts. The results underscore the superiority of our approach in achieving high-quality, text-consistent, and spatially complete 3D scene reconstructions.

#### D. Results Comparison

To objectively assess the quality of the generated images, we employ a range of evaluation metrics, including CLIP-Score, CLIP-IQA, PSNR, SSIM, and LPIPS. Table I presents the quantitative results of our evaluation. In terms of generated image quality, SceneDreamer360 consistently outperforms LucidDreamer across multiple metrics, achieving higher scores in CLIP-Score, CLIP-IQA-sharp, CLIP-IQA-colorful, and CLIP-IQA-resolution. These results indicate that our method produces sharper, more colorful, and higher-resolution images



Fig. 5. The ablation study of SceneDreamer360.

that reflect the text prompts with greater fidelity compared to LucidDreamer. Due to the significant perspective variations and incomplete scenes produced by Text2Room, we do not include it in our quantitative image quality measurements. As shown in Fig. 3, however, the scenes generated by Text2Room display poor consistency and often result in unrealistic or fragmented visual scenarios. This highlights the clear advantage of SceneDreamer360 in generating coherent and realistic multiview 3D scenes.

To evaluate the overall render quality of the 360-degree scenes, we calculate the average values for PSNR, SSIM, and LPIPS across the generated images. SceneDreamer360 consistently outperforms other methods, achieving higher PSNR and SSIM scores, which indicate superior image fidelity and structural similarity. Additionally, our method achieves lower LPIPS scores, reflecting enhanced perceptual quality in the generated images. These findings underscore the effectiveness of our approach in producing high-quality, consistent, and visually appealing 3D multiview images, thereby setting a new benchmark for text-driven 3D scene generation.

#### E. Ablation Study

We conduct ablation experiments on two critical modules—panoramic image enhancement and point cloud alignment—to evaluate their individual contributions to the effectiveness of our method. The primary configurations tested are: “Standard” (using both modules), “W/o Enhance” (without the enhancement module), “W/o Alignment” (without the alignment module), and “W/o Enhance & Alignment” (without both modules).

As illustrated in Fig. 5, the absence of the enhancement module results in a noticeable loss of fine details, indicating that this module is essential for preserving high-resolution features. When the alignment module is removed, we observe significant visual discontinuities, underscoring its role in maintaining spatial coherence across the generated scenes. Notably, the configuration without both modules produces the poorest

results, lacking both detail and spatial consistency, which detracts from the realism of the reconstructed 3D scenes.

To further quantify these effects, we measure the performance of the reconstructed images across the different configurations, as shown in Table II. The results indicate that removing the alignment module leads to a more substantial decline in metrics compared to the removal of the enhancement module, highlighting the alignment module's critical role in ensuring consistent and coherent spatial generation. These findings demonstrate that both modules are integral to our method, with alignment playing a particularly crucial role in achieving high-quality 3D scene reconstruction.

## V. CONCLUSION

In this study, we introduce a novel text-driven method, SceneDreamer360, which generates 3D panoramic scenes by leveraging large diffusion models, enabling the creation of high-quality scenes without constraints to specific target domains. First, we generate highly coherent panoramic images from input text, producing multi-view consistent, high-quality images that are seamlessly integrated into 3D space as point clouds. Next, we apply 3D Gaussian splatting to further enhance the quality of the 3D scenes. By optimizing panoramic image generation and processing, combined with 3D Gaussian splatting, SceneDreamer360 excels in generating complex 3D scenes with fine-grained details and high resolution. Extensive experiments demonstrate that SceneDreamer360 consistently produces high-quality, diverse 3D scenes under various conditions. Overall, SceneDreamer360 provides an efficient, high-quality solution for text-to-3D scene conversion, contributing to future advancements in 3D scene generation.

## REFERENCES

- [1] C. Wang, M. Chai, M. He, D. Chen, and J. Liao, “Clip-nerf: Text-and-image driven manipulation of neural radiance fields,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3835–3844.
- [2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
- [3] A. Jain, B. Mildenhall, J. T. Barron, P. Abbeel, and B. Poole, “Zero-shot text-guided object generation with dream fields,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 867–876.
- [4] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, “Dreamfusion: Text-to-3d using 2d diffusion,” *arXiv preprint arXiv:2209.14988*, 2022.
- [5] L. Höllerin, A. Cao, A. Owens, J. Johnson, and M. Nießner, “Text2room: Extracting textured 3d meshes from 2d text-to-image models,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7909–7920.
- [6] J. Chung, S. Lee, H. Nam, J. Lee, and K. M. Lee, “Luciddreamer: Domain-free generation of 3d gaussian splatting scenes,” *arXiv preprint arXiv:2311.13384*, 2023.
- [7] C. Zhang, Q. Wu, C. Cruz Gambardella, X. Huang, D. Phung, W. Ouyang, and J. Cai, “Taming stable diffusion for text to 360° panorama image generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- [8] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, “3d gaussian splatting for real-time radiance field rendering.” *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.
- [9] F. Murtagh, “Multilayer perceptrons for classification and regression,” *Neurocomputing*, vol. 2, no. 5–6, pp. 183–197, 1991.
- [10] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” *arXiv preprint arXiv:2106.09685*, 2021.
- [11] S. K. Ramakrishnan, A. Gokaslan, E. Wijmans, O. Maksymets, A. Clegg, J. Turner, E. Undersander, W. Galuba, A. Westbury, A. X. Chang *et al.*, “Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai,” *arXiv preprint arXiv:2109.08238*, 2021.
- [12] L. Zhang, A. Rao, and M. Agrawala, “Adding conditional control to text-to-image diffusion models,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.
- [13] X. Wang, L. Xie, C. Dong, and Y. Shan, “Real-esrgan: Training real-world blind super-resolution with pure synthetic data,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1905–1914.
- [14] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” 2014. [Online]. Available: <https://arxiv.org/abs/1406.2661>
- [15] E. R. Chan, M. Monteiro, P. Kellnhofer, J. Wu, and G. Wetzstein, “pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis,” 2021. [Online]. Available: <https://arxiv.org/abs/2012.00926>
- [16] M. Niemeyer and A. Geiger, “Giraffe: Representing scenes as compositional generative neural feature fields,” 2021. [Online]. Available: <https://arxiv.org/abs/2011.12100>
- [17] T. Nguyen-Phuoc, C. Li, L. Theis, C. Richardt, and Y.-L. Yang, “Hologan: Unsupervised learning of 3d representations from natural images,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7588–7597.
- [18] T. H. Nguyen-Phuoc, C. Richardt, L. Mai, Y. Yang, and N. Mitra, “Blockgan: Learning 3d object-aware scene representations from unlabelled images,” *Advances in neural information processing systems*, vol. 33, pp. 6767–6778, 2020.
- [19] P. Achlioptas, O. Diamanti, I. Mitliagkas, and L. Guibas, “Learning representations and generative models for 3d point clouds,” in *International conference on machine learning*. PMLR, 2018, pp. 40–49.
- [20] D. W. Shu, S. W. Park, and J. Kwon, “3d point cloud generative adversarial network based on tree structured graph convolutions,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3859–3868.
- [21] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models.” *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
- [22] J. Song, C. Meng, and S. Ermon, “Denoising diffusion implicit models,” *arXiv preprint arXiv:2010.02502*, 2020.
- [23] H. Hou, M. Feng, Z. Wu, W. Dong, Q. Zhu, Y. Wang, and A. Mian, “3d object detection from point cloud via voting step diffusion,” *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2024.
- [24] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, “Zero-shot text-to-image generation,” in *International conference on machine learning*. Pmlr, 2021, pp. 8821–8831.
- [25] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.
- [26] L. Zhou, Y. Du, and J. Wu, “3d shape generation and completion through point-voxel diffusion,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 5826–5835.
- [27] A. Vahdat, F. Williams, Z. Gojcic, O. Litany, S. Fidler, K. Kreis *et al.*, “Lion: Latent point diffusion models for 3d shape generation,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 10 021–10 039, 2022.
- [28] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, “Dreamfusion: Text-to-3d using 2d diffusion,” *arXiv preprint arXiv:2209.14988*, 2022.
- [29] R. Fridman, A. Abecasis, Y. Kasten, and T. Dekel, “Scenescape: Text-driven consistent scene generation,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [30] D. Cohen-Bar, E. Richardson, G. Metzer, R. Giryes, and D. Cohen-Or, “Set-the-scene: Global-local training for generating controllable nerf scenes,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2920–2929.
- [31] J. Lei, J. Tang, and K. Jia, “Rgbd2: Generative scene synthesis via incremental view inpainting using rgbd diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8422–8434.
- [32] G. Qian, J. Mai, A. Hamdi, J. Ren, A. Siarohin, B. Li, H.-Y. Lee, I. Skorokhodov, P. Wonka, S. Tulyakov *et al.*, “Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors,” *arXiv preprint arXiv:2306.17843*, 2023.

- [33] S. Zhou, Z. Fan, D. Xu, H. Chang, P. Chari, T. Bharadwaj, S. You, Z. Wang, and A. Kadambi, "Dreamscene360: Unconstrained text-to-3d scene generation with panoramic gaussian splatting," *arXiv preprint arXiv:2404.06903*, 2024.
- [34] Y. Ma, D. Zhan, and Z. Jin, "Fastscene: Text-driven fast 3d indoor scene generation via panoramic gaussian splatting," *arXiv preprint arXiv:2405.05768*, 2024.
- [35] P. Li, C. Tang, Q. Huang, and Z. Li, "Art3d: 3d gaussian splatting for text-guided artistic scenes generation," *arXiv preprint arXiv:2405.10508*, 2024.
- [36] J. Shriram, A. Trevithick, L. Liu, and R. Ramamoorthi, "Realdreamer: Text-driven 3d scene generation with inpainting and depth diffusion," *arXiv preprint arXiv:2404.07199*, 2024.
- [37] H. Zhou, X. Cheng, W. Yu, Y. Tian, and L. Yuan, "Holodreamer: Holistic 3d panoramic world generation from text descriptions," *arXiv preprint arXiv:2407.15187*, 2024.
- [38] R. Qiao, H. Ji, Z. Zhu, and W. Zhang, "Local-to-global semantic learning for multi-view 3d object detection from point cloud," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2024.
- [39] Z. Song, C. Jia, L. Yang, H. Wei, and L. Liu, "Graphalign++: An accurate feature alignment by graph matching for multi-modal 3d object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 4, pp. 2619–2632, 2024.
- [40] S. Tulsiani, H. Su, L. J. Guibas, A. A. Efros, and J. Malik, "Learning shape abstractions by assembling volumetric primitives," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2635–2643.
- [41] Z. Lu, B. Cao, and Q. Hu, "Lidar-camera continuous fusion in voxelized grid for semantic scene completion," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2024.
- [42] H. Chen, Y. Zhang, X. Wang, X. Duan, Y. Zhou, and W. Zhu, "Disendreamer: Subject-driven text-to-image generation with sample-aware disentangled tuning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 8, pp. 6860–6873, 2024.
- [43] M. Yavartanoo, J. Chung, R. Neshatavar, and K. M. Lee, "3dias: 3d shape reconstruction with implicit algebraic surfaces," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12446–12455.
- [44] T. Takikawa, J. Litalien, K. Yin, K. Kreis, C. Loop, D. Nowrouzezahrai, A. Jacobson, M. McGuire, and S. Fidler, "Neural geometric level of detail: Real-time rendering with implicit 3d shapes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11358–11367.
- [45] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.
- [46] S. Fridovich-Keil, A. Yu, M. Tancik, Q. Chen, B. Recht, and A. Kanazawa, "Plenoxels: Radiance fields without neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5501–5510.
- [47] L. Liu, J. Gu, K. Zaw Lin, T.-S. Chua, and C. Theobalt, "Neural sparse voxel fields," *Advances in Neural Information Processing Systems*, vol. 33, pp. 15651–15663, 2020.
- [48] C. Sun, M. Sun, and H.-T. Chen, "Direct voxel grid optimization: Superfast convergence for radiance fields reconstruction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5459–5469.
- [49] A. Yu, R. Li, M. Tancik, H. Li, R. Ng, and A. Kanazawa, "Plenotrees for real-time rendering of neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5752–5761.
- [50] Q. Xu, Z. Xu, J. Philip, S. Bi, Z. Shu, K. Sunkavalli, and U. Neumann, "Point-nerf: Point-based neural radiance fields," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5438–5448.
- [51] T. Müller, F. Rousselle, J. Novák, and A. Keller, "Real-time neural radiance caching for path tracing," *arXiv preprint arXiv:2106.12372*, 2021.
- [52] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM transactions on graphics (TOG)*, vol. 41, no. 4, pp. 1–15, 2022.
- [53] A. Chen, Z. Xu, A. Geiger, J. Yu, and H. Su, "Tensorf: Tensorial radiance fields," in *European conference on computer vision*. Springer, 2022, pp. 333–350.
- [54] S. J. Garbin, M. Kowalski, M. Johnson, J. Shotton, and J. Valentin, "Fastnerf: High-fidelity neural rendering at 200fps," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 14346–14355.
- [55] C. Reiser, S. Peng, Y. Liao, and A. Geiger, "Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 14335–14345.
- [56] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. De Mello, O. Gallo, L. J. Guibas, J. Tremblay, S. Khamis *et al.*, "Efficient geometry-aware 3d generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16123–16133.
- [57] Z. Chen, T. Funkhouser, P. Hedman, and A. Tagliasacchi, "Mobilenerf: Exploiting the polygon rasterization pipeline for efficient neural field rendering on mobile architectures," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16569–16578.
- [58] J. Li and M. Bansal, "Panogen: Text-conditioned panoramic environment generation for vision-and-language navigation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [59] O. Bar-Tal, L. Yariv, Y. Lipman, and T. Dekel, "Multidiffusion: Fusing diffusion paths for controlled image generation," 2023.
- [60] S. Tang, F. Zhang, J. Chen, P. Wang, and Y. Furukawa, "Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion," *arXiv*, 2023.
- [61] H. Wang, X. Xiang, Y. Fan, and J.-H. Xue, "Customizing 360-degree panoramas through text-to-image diffusion models," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 4933–4943.
- [62] S. F. Bhat, R. Birk, D. Wofk, P. Wonka, and M. Müller, "Zoedepth: Zero-shot transfer by combining relative and metric depth," *arXiv preprint arXiv:2302.12288*, 2023.
- [63] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," *arXiv preprint arXiv:2104.08718*, 2021.
- [64] J. Wang, K. C. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 2555–2563.

**Wenrui Li** received the B.S. degree from the School of Information and Software Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2021. He is currently working toward the Ph.D. degree from the School of Computer Science, Harbin Institute of Technology ( HIT ), Harbin, China. His research interests include multimedia search, joint source-channel coding, AIGC and spiking neural network.



**Fucheng Cai** is currently pursuing a Bachelor's degree in Computer Science and Technology at the School of Computer Science and Technology at Harbin Institute of Technology, under the supervision of Professor Fan Xiaopeng. His primary research interests focus on multimodal 3D scene understanding.





**Yapeng Mi** is currently working towards his Bachelor's degree in Computer Science and Technology at Harbin Institute of Technology's School of Computer Science and Technology, under the guidance of Professor Fan Xiaopeng. His main research interests lie in multimodal and 3D scene understanding.



**Xiaopeng Fan** (Senior Member, IEEE) received the B.S. and M.S. degrees from the Harbin Institute of Technology ( HIT), Harbin, China, in 2001 and 2003, respectively, and the Ph.D. degree from the Hong Kong University of Science and Technology, Hong Kong, in 2009. In 2009, he joined HIT, where he is currently a Professor. From 2003 to 2005, he was with Intel Corporation, China, as a Software Engineer. From 2011 to 2012, he was with Microsoft Research Asia, as a Visiting Researcher. From 2015 to 2016, he was with the Hong Kong University of Science and Technology, as a Research Assistant Professor. He has authored one book and more than 170 articles in refereed journals and conference proceedings. His research interests include video coding and transmission, image processing, and computer vision. He was the Program Chair of PCM2017, Chair of IEEE SGC2015, and Co-Chair of MCSN2015. He was an Associate Editor for IEEE 1857 Standard in 2012. He was the recipient of Outstanding Contributions to the Development of IEEE Standard 1857 by IEEE in 2013.



**Zhe Yang** received the B.S. degree from the College of Science, Northeast Forestry University, China, in 2022. He is currently pursuing the M.E. degree with the University of Electronic Science and Technology of China. His research interests include zero-shot learning, audio-visual learning, and computer vision.



**Wangmeng Zuo** (Senior Member, IEEE) received the Ph.D. degree in computer application technology from Harbin Institute of Technology, Harbin, China, in 2007. He is currently a Professor with the Faculty of Computing, Harbin Institute of Technology. He has published over 200 papers in top tier academic journals and conferences. His current research interests include low level vision, image/video generation, and multimodal understanding. He served as an Associate Editor for IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, and SCIENCE CHINA Information Sciences.



**Xingtao Wang** obtained his B.S. degree in Mathematics and Applied Mathematics, as well as his Ph.D. degree in Computer Science, from the Harbin Institute of Technology ( HIT) in Harbin, China, in 2016 and 2022, respectively. In 2023, he served as an Assistant Research Fellow at the School of Artificial Intelligence, HIT, and currently holds the position of Associate Researcher. His research focuses on computer graphics, digital twins, and panoramic vision.