# ShowRoom3D: Text to High-Quality 3D Room Generation Using 3D Priors

Weijia Mao[1]    Yan-Pei Cao[2*]    Jia-Wei Liu[1]    Zhongcong Xu[1]    Mike Zheng Shou[1*]

[1]Show Lab, National University of Singapore    [2]ARC Lab, Tencent PCG
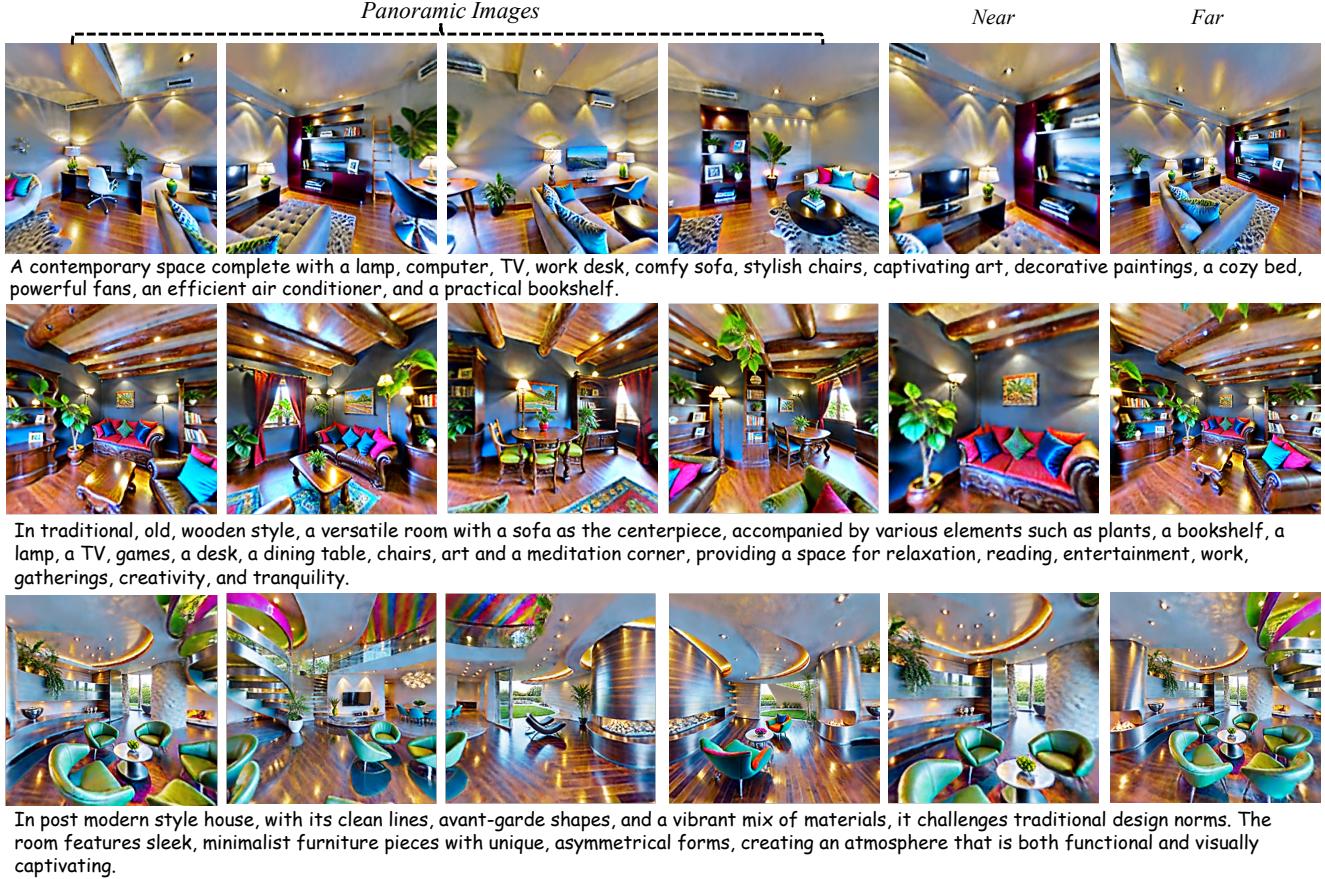
https://showroom3d.github.io

Figure 1. *ShowRoom3D:* A novel method for generating high-quality room-scale scenes that can be rendered at any position.

## Abstract

*We introduce ShowRoom3D, a three-stage approach for generating high-quality 3D room-scale scenes from texts. Previous methods using 2D diffusion priors to optimize neural radiance fields for generating room-scale scenes have shown unsatisfactory quality. This is primarily attributed to the limitations of 2D priors lacking 3D awareness and constraints in the training methodology. In this paper, we utilize a 3D diffusion prior, MVDiffusion, to optimize the 3D room-scale scene. Our contributions are in two aspects. Firstly, we propose a progressive view selection process to optimize NeRF. This involves dividing the training process into three stages, gradually expanding the camera sampling scope. Secondly, we propose the pose transformation method in the second stage. It will ensure MVDiffusion provide the*
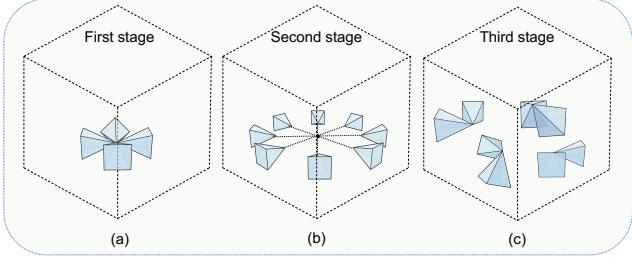
*Corresponding Author.

Figure 2. The illustration of every stage's camera sampling method. (a) In the first stage, the camera is positioned at the origin and can rotate freely. (b) In the second stage, the camera is sampled at various positions, but its direction always faces outward from the origin. (c) In the third stage, at different iterations, the camera position and perspective are randomly sampled. Within a single iteration, the cameras remain in the same position.

*accurate view guidance. As a result, ShowRoom3D enables the generation of rooms with improved structural integrity, enhanced clarity from any view, reduced content repetition, and higher consistency across different perspectives. Extensive experiments demonstrate that our method, significantly outperforms state-of-the-art approaches by a large margin in terms of user study.*

# 1. Introduction

The generation of 3D room-scale scenes is crucial for various industries, including VR/AR and Metaverse. In the 2D domain, there has been significant progress in image generation conditioned on user input, thanks to models like Stable Diffusion [32] and Imagen [33]. 2D image generation models allow users to control content using prompts and other modalities, such as layouts or poses. However, in the 3D domain, the lack of large-scale 3D datasets has led to methods [14, 26, 42, 43, 49] that combine the text-to-image diffusion model with 3D representations such as NeRF [21] or DMTet [37], often focusing on individual 3D objects generation or their combinations.

However, few of these methods tackle the challenges of 3D room-scale scene generation, as the output needs to be dense, coherent, and encompass all required structures in the views. When applied to 3D room generation, they encounter serious issues such as the Janus problem, unreasonable room structure, style inconsistencies, and more. At the same time, another line of research [8, 10] involves employing the Stable Diffusion inpainting model to generate 3D scenes. However, the consistency within indoor environments is notably subpar, characterized by visible distortions and blurring.

Recently, MVDiffusion [40] is proposed to generate the panoramic images of a 3D scene, offering several advantages. (1) It is the first model finetuned on the Stable Diffu-

sion model using the Matterport3D indoor scene dataset [3]. It will provide the model with enhanced prior knowledge about the structures and layouts of 3D rooms. (2) MVDiffusion introduces the Correspondence-Aware Attention (CAA) module to ensure consistency between views. This endows the model with 3D awareness, considering the generated images as integral parts of the entire room. However, MVDiffusion is specifically designed to generate only panoramic images of a 3D scene and cannot be used to create a fully realized 3D space. The geometry and structure of the generated scene are not guaranteed.

In our approach, we leverage MVDiffusion in conjunction with NeRF to create a 3D room-scale scene. However, optimizing a NeRF model, which accurately represents the room with high-quality geometry and appearance, using MVDiffusion is not a trivial task. There are two challenges: (1) Confirming the room's geometry and structure while also aiming for rendering from any view during the same training stage tends to yield suboptimal results. (2) The pretrained MVDiffusion model can not effectively handle scenarios where the camera is placed at any position within the room, except for the origin. In such cases, the MVDiffusion panoramic model assumes the camera is at the center of the room, providing inaccurate view guidance for NeRF training.

To address the first challenge, we adopt a progressive view selection approach and divide our training process into three distinct stages. Progressive view selection involves gradually expanding the camera sampling scope in different stages. We first ensure that the geometry of the room is well generated within a limited set of training views, and then we consider rendering the room at any position. As illustrated in Figure 2(a), during the first stage, we position the camera at the center of the room to generate a panoramic view. This initial step is crucial for determining the structure and geometry of the room. In the second stage, we continue to distill the NeRF model, ensuring that the camera consistently faces outward from the origin at any position, as depicted in Figure 2(b). This stage further improves the room's geometry and enables rendering from multiple viewpoints. In the third stage, we sample the cameras at any position and apply rotations to refine the NeRF model at different iterations. During each iteration, we randomly sample cameras from the same position, as shown in Figure 2(c). Ultimately, this process will yield a NeRF model capable of rendering the generated rooms from any position and at any rotation.

To address the second challenge, we introduce a pose transformation method to address situations where the sampled cameras are not at the origin. This transformation will provide an equivalent camera pose to the MVDiffusion model, as opposed to using the real pose. The equivalent camera will share the similar view with the real one. This

2

method ensures that MVDiffusion provides accurate view guidance, even when the sampled cameras' positions are not at the origin.

In summary, our key contributions are as follows: (1) We are the first to explore the utilization of 3D diffusion prior for generating high-quality 3D room-scale scenes using the SDS method. (2) We present a three-stage training pipeline, incorporating distinct camera sampling methods in each stage and pose transformation in the second stage to improve the clarity and aesthetics of the generated room. (3) Our method enables the generation of state-of-the-art 3D room-scale scenes, showcasing not only more compelling geometry and appearance but also a more reasonable room structure and reduced content repetition. Furthermore, it exhibits the capability to render views across a larger space, surpassing the capabilities of previous methods.

## 2. Related Work

**3D Content Generation.** The emergence of NeRF [21] has significantly improved the quality of novel view synthesis in the 3D domain. NeRF-based models [16, 18, 21, 24, 27, 39, 44] integrate volume rendering algorithms with MLPs or voxels to predict color and opacity. In the field of 3D generation, many works [2, 7, 12, 23, 35, 46] have combined 2D unconditional generative models with NeRF to create 3D contents.

DreamFusion [26] proposes score distillation sampling (SDS) to utilize a 2D text-to-image model to optimize the NeRF [21] model for generating 3D objects. However, the generated content is often oversaturated and plagued by the Janus problem (multi-head problem). Subsequent research [1, 43, 49]endeavors aimed to improve the quality of 3D content and alleviate oversaturation and the Janus problem. Prolificdreamer [43] introduced an improved method, VSD, which utilizes LoRA [11] to train with the NeRF model, effectively alleviating oversaturation. Set-the-scene [4] and CompoNeRF [15] utilize the SDS distillation method to create scenes with straightforward object compositions. However, these approaches are limited to generating very basic scenes comprising only a few objects. Several studies [17, 20, 36, 50] also employ the SDS-based approaches to edit the NeRF conditioned on users' prompts.

Several work has developed 3D generation models based on 2D generation models using 3D datasets. Zero123 [19] uses the Objvarse [6] dataset to finetune the Stable Diffusion model [32], enabling it to generate view-consistent images. Magic123 [28] leverages both the Stable Diffusion model and the Zero123 model as 2D and 3D priors to optimize NeRF, resulting in a consistent structure for 3D objects. MVDream [38] redesigns the Stable Diffusion architecture, incorporating a 3D-aware attention module. It uses the Objvarse dataset [6] for finetuning to generate high-quality content that mitigates the Janus problem. However,

these efforts predominantly center around 3D object generation, with limited focus on generating high-quality indoor scenes.

**Indoor Scene Generation.** Historically, several works [13, 30, 31, 41, 45] have utilized real indoor scene datasets like Matterport3D [3], ScanNet [5], and RealEstate [48] datasets to train generative models, such as GANs or autoregressive transformers, for synthesizing novel views. However, these approaches have primarily focused on generating novel views within indoor scenes, without capturing the entire 3D indoor environment. As a result, the generated content quality has been suboptimal in consistency. At the same time, Scenescape [8], Text2room [10] and Text2NeRF [47] utilize the inpainting function of Stable Diffusion models to generate 3D scenes. Nevertheless, these methods still face challenges in maintaining style and content consistency, as the generation of each image is conditioned solely on the previous image.

MVDiffusion [40] aims to address this issue by introducing a correspondence attention module to capture relationships between views. The camera poses in the generated views will be fed into MVDiffusion model. Besides, it finetunes the Stable Diffusion model using the Matterport3D [3] dataset. However, it only can generate view-consistent panoramic images. There are limitations in assuring the geometry of the room and rendering the room from any view.

## 3. Method

In this section, we first briefly introduce the score distillation sampling (SDS) method and the MVDiffusion model which our method is based on (Section. 3.1). Then we will introduce our task setting and our approach ShowRoom3D that adopts three-stage training procedure (Section. 3.2). Our pipeline is shown in Figure 3.

### 3.1. Preliminary

**Text-to-3D Generation by Score Distillation Sampling (SDS) [26].** SDS is a promising method that combines 2D generative models, such as Stable Diffusion [32], with 3D representations, like Neural Radiance Fields (NeRF) [21], to generate 3D objects and scenes solely based on textual prompts. When using a pretrained text-to-image model, noise is introduced into the NeRF's output. Then the noisy image $\mathbf{x}_t$, a time step $t$, a text embedding $y$ are then fed into a U-Net architecture to predict the noise, denoted as $\hat{\epsilon}$. The optimization of NeRF's parameters, denoted as $\phi$, is based on minimizing the loss between the noise and the predicted noise. In recent times, distillation methods have been effectively applied in the field of 3D generation, producing
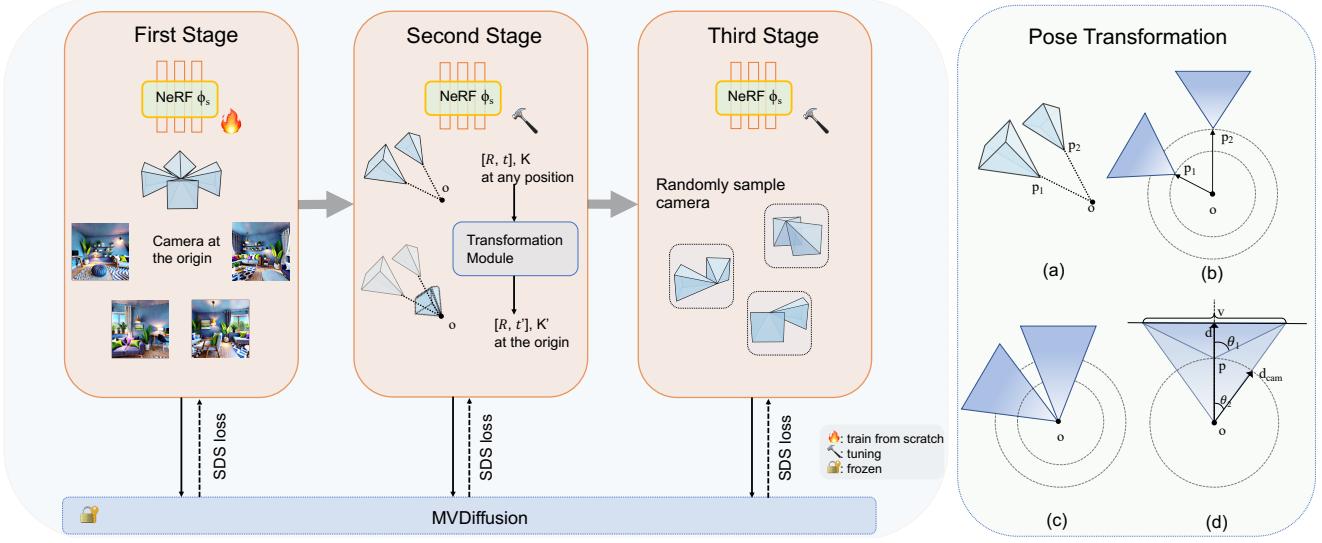
Figure 3. Method overview: **Left**: Our three-stage training pipeline. First Stage: the camera will be at the center of the room and rotate any degree. Second stage: the camera will be at any position and face outward from the center o. Third stage: the camera will be at any position and rotate any degree. **Right**: It introduces the pose transformation module in the second stage. (a) shows the camera sampling method in the second stage. p1 and p2 are sampled points at one iteration and o is the center of the room in 3D space. (b) shows the perspective of (a) observed from a 2D plane. (c) shows the two new cameras after the pose transformation. (d) shows the specific scenes observed by the camera at each sampling point. v, $d_{\text{cam}}$, d represent the observed view, the distance from the camera position to the origin and the depth of the view. $\theta_1$ and $\theta_2$ represent the FOV of two cameras.

significant improvements. Its gradient is approximated by

$$\nabla_\phi \mathcal{L}_{\text{SDS}}(\phi) \approx \mathbb{E}_{t,\epsilon,c} \left[ \omega(t)(\hat{\epsilon}(\mathbf{x}_t, t, y) - \epsilon) \frac{\partial \mathbf{g}(\phi, c)}{\partial \phi} \right], \tag{1}$$

where $\omega(t)$ is a weighting function. Despite SDS has made significant strides in 3D object generation, how to use the SDS method specifically for room-scale scene generation remains underexplored.

**MVDiffusion [40].** The MVDiffusion model is a multi-view consistency generation model built upon Stable Diffusion. One of the crucial components that ensures view consistency is the correspondence-aware attention mechanism (CAA module), which is employed to capture the relationship between adjacent views. This mechanism takes into account both the source feature maps, denoted as $F$, and the target feature maps, denoted as $F^l$.

For each source feature, MVDiffusion leverages the relative camera pose to determine the location of the corresponding target feature. $s$ and $t$ represent the source and target pixel. $\bar{F}(\mathbf{s})$ is the source feature with the positional encoding. $l$ is the number of target feature maps and $\bar{F}^l(t_*^l)$ is the target feature with the positional encoding. $\mathcal{N}(\mathbf{t}^l)$ is the neighborhood of the target pixel. Subsequently, the output of the attention mechanism is calculated as follows,

$$\begin{aligned}
\mathbf{Q} &= \mathbf{W_Q}\bar{F}(\mathbf{s}), \\
\mathbf{K} &= \mathbf{W_K}\bar{F}^l(t_*^l), \\
\mathbf{m} &= \sum_l \sum_{t_*^l \in \mathcal{N}(\mathbf{t}^l)} \text{SoftMax}(\mathbf{Q} \cdot \mathbf{K}) \cdot \mathbf{W_V}\bar{F}^l(t_*^l),
\end{aligned} \tag{2}$$

where $\mathbf{W_Q}$, $\mathbf{W_V}$, and $\mathbf{W_K}$ represent the query, value, and key components of the attention mechanism. For clarity, we do not introduce some specific details, such as position encoding and adding integer displacements. When we use the MVDiffusion model to generate the panoramic images, the prompt and the camera poses of images need to be fed into it. The camera poses will be used in the CAA module to calculate the relationship between the source feature and the target features.

### 3.2. ShowRoom3D

**Task Setting.** In this paper, we propose a new method to generate the high-quality room-scale scene that can be rendered at any position. Given by text prompts $y$, our objective is to generate a 3D room represented by a NeRF model $\Phi$ using the pretrained MVDiffusion model $M$. We choose Instant-NGP [22] as our NeRF representation due to its numerous advantages, including rapid coverage speed and the ability to reconstruct complex geometries.

**First Stage.** As the MVDiffusion model initially generates a panoramic scene, we employ the SDS method to optimize

a NeRF model, specifically designed to represent a room-scale scene panorama. Initially, we position the camera at the world coordinate center and randomly sample the rotation degree to obtain the camera pose like in Figure 2(a). This camera pose is then fed into both the NeRF model and the correspondence-aware attention module of the MVDiffusion model. So the gradient can be approximated by

$$\nabla_\phi \mathcal{L}(\phi) \approx \mathbb{E}_{t,\epsilon,c} \left[ \omega(t)(\hat{\epsilon}(\mathbf{x}_t, t, y, c) - \epsilon) \frac{\partial \mathbf{g}(\phi, c)}{\partial \phi} \right]. \tag{3}$$

In this equation, $y$ represents the prompt, and $c$ represents the camera pose, including the camera's extrinsics and intrinsics. We will input $c$ into the NeRF model to obtain the rendered image. Subsequently, we introduce noise to the rendered image, resulting in the noisy latent $\mathbf{x}_t$. This noisy latent $\mathbf{x}_t$, the prompt $y$, the camera pose $c$ and the timestep $t$ are then fed into the pretrained MVDiffusion model's U-Net module and CAA module for noise prediction and gradient calculation.

**Second Stage.** After the first stage, we obtain the panoramic NeRF, which initially determines the geometry and layout of the room. We can rotate any degree to render this room, however, it restricts the camera's position, preventing us from rendering the room from any position. At the same time, the geometry of the room is also subpar due to the limited training views.

In the second stage, our goal has two aspects. Firstly, this stage serves to improve the geometry and room layout by adding training views. Secondly, we can render the room in a larger space compared to the first stage. We modify the camera sampling method, as shown in the Figure 2(b). Now, the camera can be sampled at any position, and it always faces outward from the origin.

The challenge lies in ensuring that MVDiffusion offers precise view guidance even when cameras are not positioned identically. To solve this problem, we propose pose transformation to obtain a new camera pose to be fed into the MVDiffusion model, rather than using the real camera pose directly. For any sampled camera not positioned at the center, we will employ pose transformation to obtain an equivalent new pose at the center. The core idea is that the new camera pose at the center will have a smaller field of view (FOV). Even if the new camera is farther from the scene, the smaller FOV will help it see the similar view compared with the old camera. Next, we will demonstrate the procedure for calculating the new camera pose.

In Figure 3(a)(b), $p1$ and $p2$ are two points we sampled at one iteration, which represent the camera positions. In Figure 3(d), let point $o$ be the center of the room and the world coordinate, and point $p$ represent the camera position. $\theta_1$ denotes the FOV of the camera at position $p$. $v$ represents the region visible to the camera and $d$ is the averaged depth of this region. The real camera is located at $p$.

Pose transformation is used to calculate the new camera at the position $o$. $\theta_2$ is the FOV of the new camera at position $o$. To ensure that two cameras have a similar view $v$, we calculate $\theta_2$ using the following formulas.
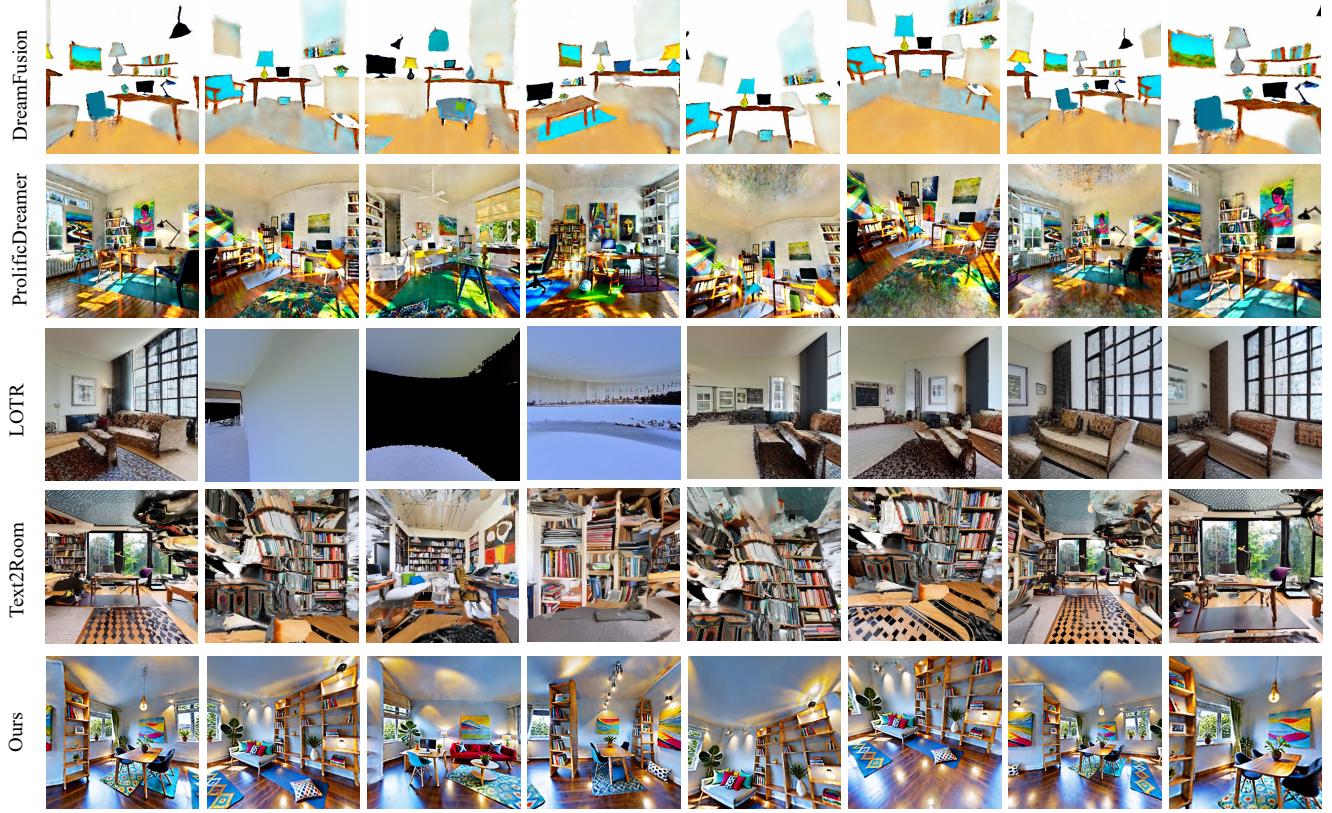
$$\begin{aligned} \frac{v}{2} &= \tan \theta_2 \cdot d, \\ \frac{v}{2} &= \tan \theta_1 \cdot (d - d_{\text{cam}}), \\ \theta_2 &= \arctan \left( \frac{\tan \theta_1 \cdot (d - d_{\text{cam}})}{d} \right), \end{aligned} \tag{4}$$

where $d_{\text{cam}}$, $\theta_1$ are known parameters and the depth $d$ can be obtained by the first stage's prior. So we can calculate the new $\theta_2$. Then we can use the rendered view from the camera at position $o$ approximate the view from the camera at position $p$ by changing the FOV $\theta_1$ to $\theta_2$.

After the pose transformation, we obtain two different yet similar camera poses $c(R, t, K)$, which represents the real camera pose fed into the NeRF model, and $c'(R, t', K')$, an approximated new camera pose fed into MVDiffusion. The rotation matrix $R$ remains constant, while the camera position $t$ and the intrinsics $K$ undergo changes. Then we can confirm that even if the sampled points are at different positions $p1$ and $p2$, the new cameras' positions fed into MVDiffusion are the same position $o$. This ensures that MVDiffusion provides the most accurate guidance about the relationships between views.

**Third Stage.** After the two-stage training process, we have established the geometry and structure of the room, enabling rendering from a wide range of perspectives. However, some issues remain after the two-stage training. The first concern is the approximated nature of the pose transformation method, which may not yield highly accurate results. The depth prior is not accurate and the rendering views from the real camera pose and the approximated new camera pose are not perfectly identical. Furthermore, during the first two stages of training, the camera consistently faces outward from the origin, resulting in some missing training views.

In the third stage, we position the camera freely and applied various degrees of rotation to further finetune the NeRF model. Following this stage, the NeRF becomes a versatile renderer capable of handling scenes from any position and at any rotation. To be specific, as shown in the Figure 2(c), we sample two points at the same position at one iteration fed into NeRF and MVDiffusion. It will ensure MVDiffusion provide the relatively accurate guidance when the sampled cameras are the same position, even if they are not at the center. At different iterations, the camera position and perspective are randomly sampled.

5

Figure 4. Qualitative comparisons of ShowRoom3D and state-of-the-art approaches.
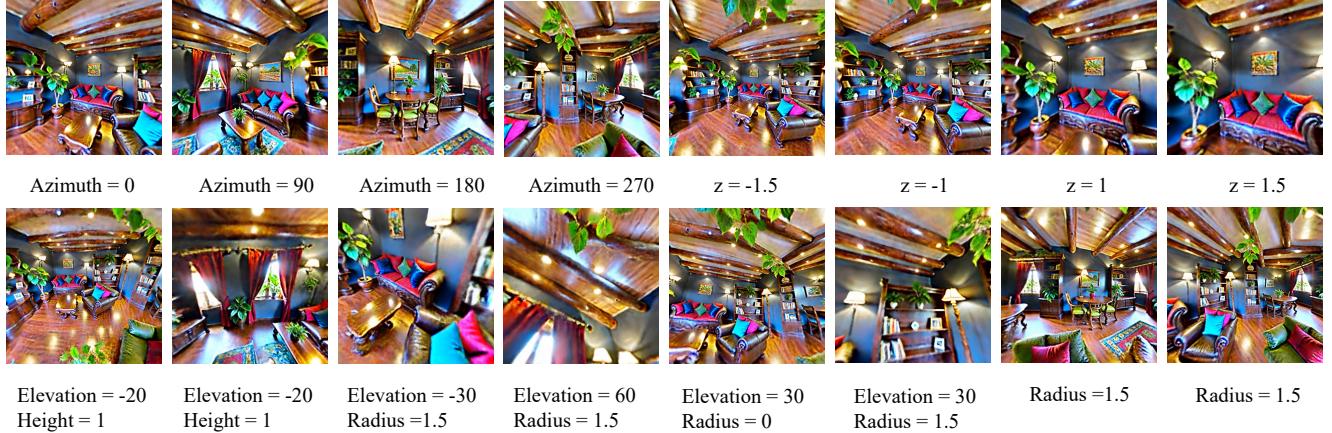
## 4. Experiments

### 4.1. Implemention details

In our approach, we utilize the panoramic model from MVDiffusion [40] as the foundational model for optimizing the NeRF [21]. We maintain the same world coordinate system and camera coordinate system as MVDiffusion. We implement the SDS method in the threestudio framework [9]. The first stage of NeRF training requires 10,000 iterations, the second stage needs 15,000 iterations and the third stage needs 5000 iterations. We utilize a single NVIDIA RTX 3090 GPU to train the three-stage NeRF model. To mitigate the issue of oversaturation introduced by the SDS method, we draw inspiration from MVDream [38] and employ similar techniques. Specifically, we anneal the timestep to control the noise added to the NeRF's output. Another strategy involves using negative prompts to guide the training process. More details will be provided in the supplementary materials.

### 4.2. Baselines

We select four state-of-the-art works to compare with our method. (1) *DreamFusion* [26] is the first work to use the text-to-image diffusion model to optimize a 3D object. (2) *ProlificDreamer* [43] introduces a new method, VSD, for optimizing a NeRF. VSD combines the vanilla Stable Diffusion model with the LORA [11] model to jointly optimize a NeRF, resulting in the alleviation of oversaturation phenomena. (3) *Text2Room* [10] is another method that leverages a 2D diffusion model to generate 3D scenes. It utilizes the Stable Diffusion inpainting model to generate 2D images sequentially. (4) *Look Outside the Room* [30] is a method that generates novel view images based on previously generated images and camera poses. It trains an autoregressive transformer model from scratch using the Matterport3D [3] dataset and RealEstate [48] dataset.

### 4.3. Qualitative Results

In Figure 11, we present RGB rendering images of scenes with the crucial coordinate information for our method and

6

| Azimuth = 0 | Azimuth = 90 | Azimuth = 180 | Azimuth = 270 | z = -1.5 | z = -1 | z = 1 | z = 1.5 |

| Elevation = -20 Height = 1 | Elevation = -20 Height = 1 | Elevation = -30 Radius =1.5 | Elevation = 60 Radius = 1.5 | Elevation = 30 Radius = 0 | Elevation = 30 Radius = 1.5 | Radius =1.5 | Radius = 1.5 |

In traditional, old, wooden style, a versatile room with a sofa as the centerpiece, accompanied by various elements such as plants, a bookshelf, a lamp, a TV, games, a desk, a dining table, chairs, art and a meditation corner, providing a space for relaxation, reading, entertainment, work, gatherings, creativity, and tranquility.

Figure 5. More views of our result.

| | Metrics | | Human Preference | | |
| --- | --- | --- | --- | --- | --- |
| | CLIP Score (↑) | Aesthetic Score (↑) | Textual Alignment (↑) | Consistency (↑) | Overall Quality (↑) |
| DreamFusion [26] | 23.56 | 4.65 | 2.75 | 2.80 | 2.55 |
| ProlificDreamer [43] | 22.45 | 4.98 | 3.10 | 3.12 | 2.87 |
| Text2Room [10] | 20.41 | 5.21 | 3.34 | 2.97 | 3.20 |
| LOTR [30] | 13.12 | 4.23 | 1.25 | 1.21 | 1.78 |
| ShowRoom3D (Ours) | **25.62** | **5.56** | **4.55** | **4.89** | **4.59** |

Table 1. Quantitative comparisons of ShowRoom3D against state-of-the-art approaches.

baselines. The initial four images depict a panoramic view of the entire room, followed by four additional perspectives in the subsequent set of images. DreamFusion and ProlificDreamer struggle to generate the correct room structure, exhibiting issues such as the Janus problem and style inconsistency. This results in blurriness in certain views. Text2Room results in view inconsistencies, stretching, and blurring due to its inpainting method. Meanwhile, LOTR fails to create panoramic scenes and exhibits poor content consistency between frames. Additionally, we present additional views of the rooms generated by our results in Figure 5. Displaying 16 views of a room, it demonstrates that our method produces not only high-quality rooms but can also render in a larger space. We have briefly annotated the crucial coordinate information, and further results comparing baselines with our method can be found in the supplementary material.
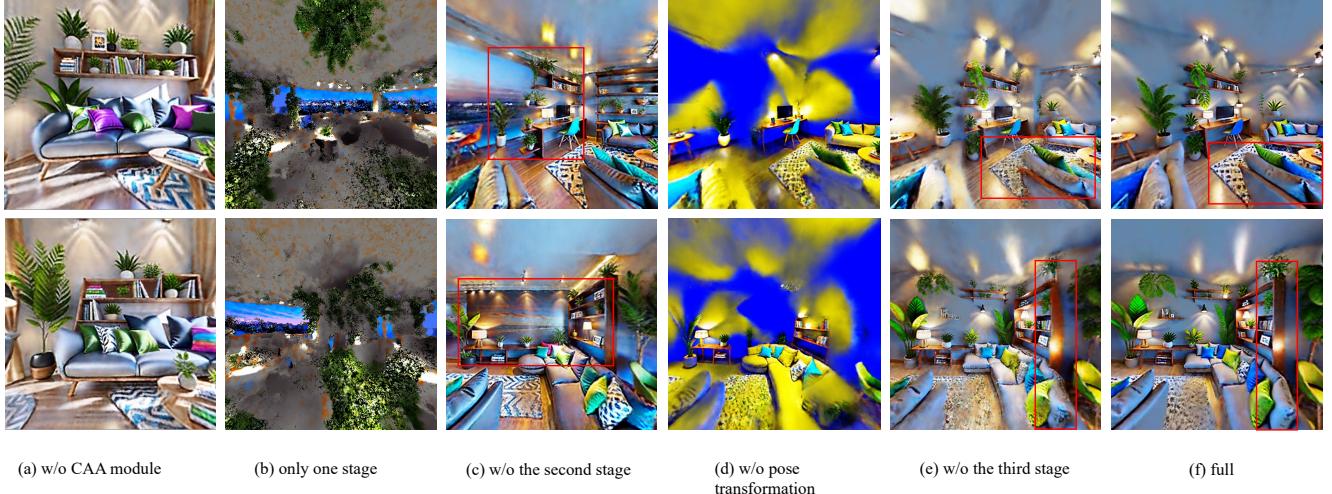
## 4.4. Quantitative Results

We compute the CLIP Score [29] and aesthetic score [34] for 120 RGB renderings of each scene and conduct a user study for scene evaluation. The CLIP score is employed to assess the alignment between rendered views and provided prompts. Additionally, the aesthetic score introduced by LAION measures the aesthetic quality of the generated images. Recent methods [25] have demonstrated its authenticity, surpassing the reliability of FID. We present quantitative results, averaged across multiple scenes, in Table 1. In this table, we observe that our method achieves the highest averaged CLIP score which means our method can generate scenes that closely align with user prompts and the highest averaged aesthetic score which means our results maintain the highest quality.

Additionally, in our user study, we leverage Amazon MTurk [1] to recruit 24 participants to rank the results obtained from our methods and other baselines. The rendered views from the generated 3D scene are evaluated across three aspects: overall quality, text alignment, and style consistency. Scores are calculated with a rating of 5 for the best-ordered view and 1 for the last. In the end, we gather a total of 259 data points to calculate the final scores. As illustrated in Table 1, we outperform other methods by a substantial margin, indicating our superior performance in terms of overall quality, text alignment, and consistency.

---
[1] https://requester.mturk.com/

| (a) w/o CAA module | (b) only one stage | (c) w/o the second stage | (d) w/o pose transformation | (e) w/o the third stage | (f) full |

In realistic style, a versatile room with a sofa as the centerpiece, accompanied by various elements such as plants, a bookshelf, a lamp, a TV, games, a desk, a dining table, chairs, art and a meditation corner, providing a space for relaxation, reading, entertainment, work, gatherings, creativity, and tranquility.

Figure 6. Ablation study on each proposed component. We choose two different views of a room. (a) We use the SD(finetuned on Matterport3D) model without CAA module to follow our three-stage pipeline. The quality of the room is also high but there are many repetitive contents in different views. Results in (b), (c), (d), and (e) demonstrate the impact of removing each corresponding component, showcasing varying degrees of quality degradation. Finally, (f) presents the result of our full method.

|  | Metrics | |
|---|---|---|
|  | CLIP Score (↑) | Aesthetic Score (↑) |
| Ours w/o CAA module | 28.23 | **5.62** |
| Ours only one stage | 14.48 | 4.91 |
| Ours w/o the second stage | 27.94 | 5.21 |
| Ours w/o pose transformation | 27.63 | 5.31 |
| Ours w/o the third stage | 26.43 | 5.44 |
| ShowRoom3D (Ours full) | **28.82** | 5.59 |

Table 2. Quantitative ablations of ShowRoom3D.

## 4.5. Ablations

Our method's key components include three-stage training and pose transformation in the second stage. We also calculate CLIP Score and aesthetic score of every ablation part, as shown in Table 2. Our method get the highest CLIP Score and the second highest aesthetic score. Next we will illustrate each component.

**The Effect of CAA Module.** We just use the Stable Diffusion(finetuned on the Matterport3D dataset) without CAA module as a prior to follow our three-stage pipeline. In Figure 6(a), we investigate the impact of the correspondence attention module. It is evident that the use of the correspondence attention module mitigates the Janus problem and improves content diversity of the room in our method, as shown in Figure 6(f).

The scores of our method are similar to those without the CAA module, proving the utility of our three-stage training pipeline for both MVDiffusion and Stable Diffusion. While our method's aesthetic score is slightly lower than without

the CAA module, this discrepancy may be attributed to the limitations of the two metrics used, as they cannot effectively assess the Janus problem.

**The Effect of Training with Only One Stage.** In Figure 6(b), we investigate the impact of three-stage training. It is evident that if we employ the SDS method to optimize the NeRF with only one stage, it struggles to generate meaningful content. This is primarily due to the CAA module's inability to correctly handle the camera pose, resulting in failed training.

**The Effect of the Second Stage.** As depicted in Figure 6(c), if we train the NeRF model without the second stage and proceed directly to the third stage after the first stage's training, the geometric quality deteriorates. Some furniture pieces may not be generated effectively, resulting in suboptimal wall and shelf shapes, as illustrated in Figure 6(c).

**The Effect of Pose Tranformation.** In the Figure 6(d), we explore the consequences of further NeRF optimization without pose transformation. The results indicate that continued NeRF optimization can lead to training instability and sometimes the training will fail because the noise from MVDiffusion model will not provide the accurate view guidance.

**The Effect of the Third Stage.** As depicted in Figure 6(e), we investigate the significance of the third stage. Without the third training stage, the quality of rendered views in this room deteriorates because in the first two stages, the camera poses are not randomly sampled, resulting in some missing

views that are left uncovered.

## 5. Conclusion

In this work, we introduce ShowRoom3D, a three-stage pipeline using a 3D diffusion prior, MVDiffusion, to optimize NeRF for the generation of high-quality 3D room-scale scenes. We employ progressive view selection approach in three stages. During the second stage, we utilize pose transformation to ensure accurate guidance from MVDiffusion. As a result, we can produce high-quality room-scale scenes which can be rendered at any position.

**Limitations.** Our approach enables the generation of high-quality 3D room-scale scenes from texts. However, there are some limitations. Firstly, similar to previous methods, our approach produces oversaturated results due to the SDS loss, despite employing certain training techniques to alleviate this occurrence. Secondly, our method is time-consuming due to the three-stage training process.

**Acknowledgement.** We thank Ziteng Gao, Hai Ci and Yiquan Chen for their helpful discussions.

## References

[1] Mohammadreza Armandpour, Ali Sadeghian, Huangjie Zheng, Amir Sadeghian, and Mingyuan Zhou. Re-imagine the negative prompt algorithm: Transform 2d diffusion into 3d, alleviate janus problem and beyond. *CoRR*, abs/2304.04968, 2023. 3

[2] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, pages 16102–16112. IEEE, 2022. 3

[3] Angel X. Chang, Angela Dai, Thomas A. Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from RGB-D data in indoor environments. In *3DV*, pages 667–676. IEEE Computer Society, 2017. 2, 3, 6, 14

[4] Dana Cohen-Bar, Elad Richardson, Gal Metzer, Raja Giryes, and Daniel Cohen-Or. Set-the-scene: Global-local training for generating controllable nerf scenes. *CoRR*, abs/2303.13450, 2023. 3

[5] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 2432–2443. IEEE Computer Society, 2017. 3

[6] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *CVPR*, pages 13142–13153. IEEE, 2023. 3

[7] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. GRAM: generative radiance manifolds for 3d-aware image generation. In *CVPR*, pages 10663–10673. IEEE, 2022. 3

[8] Rafail Fridman, Amit Abecasis, Yoni Kasten, and Tali Dekel. Scenescape: Text-driven consistent scene generation. *CoRR*, abs/2302.01133, 2023. 2, 3

[9] Yuan-Chen Guo, Ying-Tian Liu, Ruizhi Shao, Christian Laforte, Vikram Voleti, Guan Luo, Chia-Hao Chen, Zi-Xin Zou, Chen Wang, Yan-Pei Cao, and Song-Hai Zhang. threestudio: A unified framework for 3d content generation. https://github.com/threestudio-project/threestudio, 2023. 6

[10] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. *CoRR*, abs/2303.11989, 2023. 2, 3, 6, 7, 12

[11] 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. In *ICLR*. OpenReview.net, 2022. 3, 6, 12

[12] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 867–876, 2022. 3

[13] Jing Yu Koh, Harsh Agrawal, Dhruv Batra, Richard Tucker, Austin Waters, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. Simple and effective synthesis of indoor 3d scenes. In *AAAI*, pages 1169–1178. AAAI Press, 2023. 3

[14] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *CVPR*, pages 300–309. IEEE, 2023. 2

[15] Yiqi Lin, Haotian Bai, Sijia Li, Haonan Lu, Xiaodong Lin, Hui Xiong, and Lin Wang. Componerf: Text-guided multi-object compositional nerf with editable 3d scene layout. *CoRR*, abs/2303.13843, 2023. 3

[16] Jia-Wei Liu, Yan-Pei Cao, Weijia Mao, Wenqiao Zhang, David Junhao Zhang, Jussi Keppo, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Devrf: Fast deformable voxel radiance fields for dynamic scenes. *Advances in Neural Information Processing Systems*, 35:36762–36775, 2022. 3

[17] Jia-Wei Liu, Yan-Pei Cao, Jay Zhangjie Wu, Weijia Mao, Yuchao Gu, Rui Zhao, Jussi Keppo, Ying Shan, and Mike Zheng Shou. Dynvideo-e: Harnessing dynamic nerf for large-scale motion-and view-change human-centric video editing. *arXiv preprint arXiv:2310.10624*, 2023. 3

[18] Jia-Wei Liu, Yan-Pei Cao, Tianyuan Yang, Zhongcong Xu, Jussi Keppo, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Hosnerf: Dynamic human-object-scene neural radiance fields from a single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18483–18494, 2023. 3

[19] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. *CoRR*, abs/2303.11328, 2023. 3

[20] Aryan Mikaeili, Or Perel, Daniel Cohen-Or, and Ali Mahdavi-Amiri. SKED: sketch-guided text-based 3d editing. *CoRR*, abs/2303.10735, 2023. 3

[21] 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 (1)*, pages 405–421. Springer, 2020. 2, 3, 6, 12

[22] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 4

[23] Michael Niemeyer and Andreas Geiger. GIRAFFE: representing scenes as compositional generative neural feature fields. In *CVPR*, pages 11453–11464. Computer Vision Foundation / IEEE, 2021. 3

[24] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *ICCV*, pages 5845–5854. IEEE, 2021. 3

[25] 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. *CoRR*, abs/2307.01952, 2023. 7

[26] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*. OpenReview.net, 2023. 2, 3, 6, 7, 12

[27] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *CVPR*, pages 10318–10327. Computer Vision Foundation / IEEE, 2021. 3

[28] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, and Bernard Ghanem. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *CoRR*, abs/2306.17843, 2023. 3

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 7

[30] Xuanchi Ren and Xiaolong Wang. Look outside the room: Synthesizing A consistent long-term 3d scene video from A single image. In *CVPR*, pages 3553–3563. IEEE, 2022. 3, 6, 7, 12

[31] Robin Rombach, Patrick Esser, and Björn Ommer. Geometry-free view synthesis: Transformers and no 3d priors. In *ICCV*, pages 14336–14346. IEEE, 2021. 3

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10674–10685. IEEE, 2022. 2, 3, 12

[33] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 2

[34] Christoph Schuhmann. Clip+mlp aesthetic score predictor. https://github.com/christophschuhmann/improved-aesthetic-predictor, 2023. 7

[35] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. GRAF: generative radiance fields for 3d-aware image synthesis. In *NeurIPS*, 2020. 3

[36] Etai Sella, Gal Fiebelman, Peter Hedman, and Hadar Averbuch-Elor. Vox-e: Text-guided voxel editing of 3d objects. *CoRR*, abs/2303.12048, 2023. 3

[37] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In *NeurIPS*, pages 6087–6101, 2021. 2

[38] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *CoRR*, abs/2308.16512, 2023. 3, 6

[39] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *CVPR*, pages 5449–5459. IEEE, 2022. 3

[40] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *CoRR*, abs/2307.01097, 2023. 2, 3, 4, 6, 12

[41] Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In *CVPR*, pages 16773–16783. IEEE, 2023. 3

[42] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *CVPR*, pages 12619–12629. IEEE, 2023. 2

[43] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *CoRR*, abs/2305.16213, 2023. 2, 3, 6, 7, 12

[44] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, pages 16210–16220, 2022. 3

[45] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *CVPR*, pages 7465–7475. Computer Vision Foundation / IEEE, 2020. 3

[46] Eric Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Wenqing Zhang, Song Bai, Jiashi Feng, and Mike Zheng Shou. PV3D: A 3d generative model for portrait video generation. In *ICLR*. OpenReview.net, 2023. 3

[47] Jingbo Zhang, Xiaoyu Li, Ziyu Wan, Can Wang, and Jing Liao. Text2nerf: Text-driven 3d scene generation with neural radiance fields. *CoRR*, abs/2305.11588, 2023. 3

[48] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Trans. Graph.*, 37(4): 65, 2018. 3, 6

[49] Junzhe Zhu and Peiye Zhuang. Hifa: High-fidelity text-to-3d with advanced diffusion guidance. *CoRR*, abs/2305.18766, 2023. 2, 3

[50] Jingyu Zhuang, Chen Wang, Lingjie Liu, Liang Lin, and Guanbin Li. Dreameditor: Text-driven 3d scene editing with neural fields. *CoRR*, abs/2306.13455, 2023. 3

# ShowRoom3D: Text to High-Quality 3D Room Generation Using 3D Priors

## Supplementary Material

This supplementary mainly includes the implementation details of ShowRoom3D and other baselines, more comparisions of ShowRoom3D against other baselines and more ablation results.

Furthermore, we also provide a **supplementary video** to show 360° free-viewpoint renderings of the rooms from ShowRoom3D and the comparisons of ShowRoom3D against baselines.

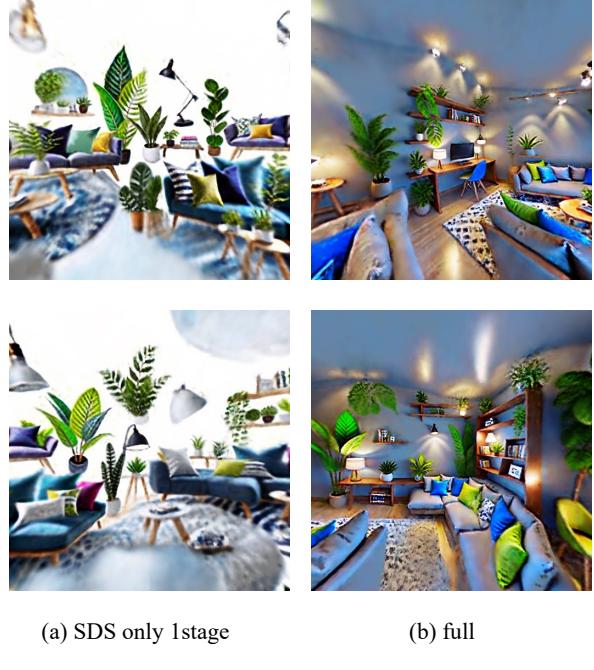## A. More Implementation Details of ShowRoom3D

We align our coordinates with MVDiffusion [40] by adopting the same right-handed world coordinate system and camera coordinate system. Specifically, the x-axis faces forward, the z-axis faces right, and the y-axis faces up. We also take the annealing time tragedy to alleviate the oversaturation phenomenon. In the first stage, the maximum time step is reduced from 0.6 to 0.02, while the minimum timestep is adjusted from 0.98 to 0.7. In the second stage, for the first 10,000 iterations, the maximum and minimum timestep values are set to 0.7 and 0.02, respectively. For the subsequent 5,000 iterations and during the third stage, the maximum and minimum timestep values become 0.4 and 0.02. We also utilize negative prompts to guide the optimization of NeRF. The negative prompt includes descriptors such as 'ugly, bad anatomy, blurry, pixelated, obscure, unnatural colors, poor lighting, dull, unclear, cropped, lowres, low quality, artifacts, duplicate, morbid, mutilated, poorly drawn face, deformed, dehydrated, bad proportions.'

## B. More Implementation Details of Baselines

**DreamFusion [26] and ProlificDreamer [43].** To ensure a fair comparison with our method, we adopt the same right-handed world coordinate system and align the camera coordinate system accordingly. In this setup, the x-axis points forward, the z-axis points right, and the y-axis points up. While DreamFusion initially operates in an object-centered camera system for 3D object generation, we modify it to a camera-centered system for consistency.

**Text2Room [10].** We adhere to the Text2Room pipeline to generate the room mesh. Subsequently, we employ Poisson surface reconstruction in Meshlab, instead of Python, for increased efficiency in rendering the mesh and comparing the rendered images with our results.

**Look Outside The Room(LOTR) [30].** Because the LOTR do not generate the novel view images according to the user's input, we use Stable Diffusion to generate the first



(a) SDS only 1stage          (b) full

In realistic style, a versatile room with a sofa as the centerpiece, accompanied by various elements such as plants, a bookshelf, a lamp, a TV, games, a desk, a dining table, chairs, art and a meditation corner, providing a space for relaxation, reading, entertainment, work, gatherings, creativity, and tranquility.

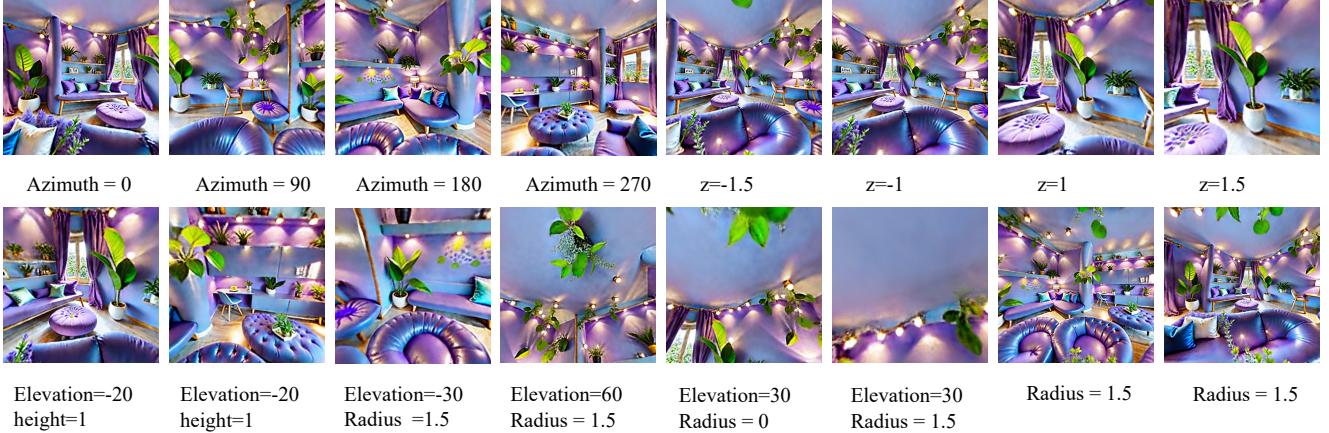Figure 7. Ablation study on one more component.

image and use the first image and the corresponding camera poses to generate the following images.

## C. More Comparisons with Baselines

We can see the results in Figure 11. We also compare the training time and inference time of our method with other baselines, as depicted in the table.

**DreamFusion.** DreamFusion utilizes the vanilla Stable Diffusion [32] to distill NeRF [21] in a single stage, leading to several disadvantages. Firstly, the model lacks prior knowledge about 3D indoor scenes, resulting in a severe Janus problem during room generation. As illustrated in Figure 11, there are numerous repetitive contents between views, creating a disjointed appearance that does not resemble a cohesive indoor scene but rather a combination of 2D images. Secondly, due to training in a single stage, the room structure is improper, with furniture and the ceiling not consistently placed at the same level.

**ProlificDreamer.** ProlificDreamer shares the same disadvantages with DreamFusion. ProlificDreamer incorporates the LORA [11] model for joint training with NeRF to ad-

| Azimuth = 0 | Azimuth = 90 | Azimuth = 180 | Azimuth = 270 | z=-1.5 | z=-1 | z=1 | z=1.5 |

| Elevation=-20 height=1 | Elevation=-20 height=1 | Elevation=-30 Radius =1.5 | Elevation=60 Radius = 1.5 | Elevation=30 Radius = 0 | Elevation=30 Radius = 1.5 | Radius = 1.5 | Radius = 1.5 |

In lavender style, a versatile room with a sofa as the centerpiece, accompanied by various elements such as plants, a bookshelf, a lamp, a TV, games, a desk, a dining table, chairs, art and a meditation corner, providing a space for relaxation, reading, entertainment, work, gatherings, creativity, and tranquility.

Figure 8. Our more results with 16 views

|  | DreamFusion | ProlificDreamer | Text2Room | LOTR | ShowRoom3D |
|---|---|---|---|---|---|
| Training Time | 1h30min | 7h20min | 2h | –– | 9h30min |
| Inference Time | 26s | 31s | 3min17s | 10min | 27s |

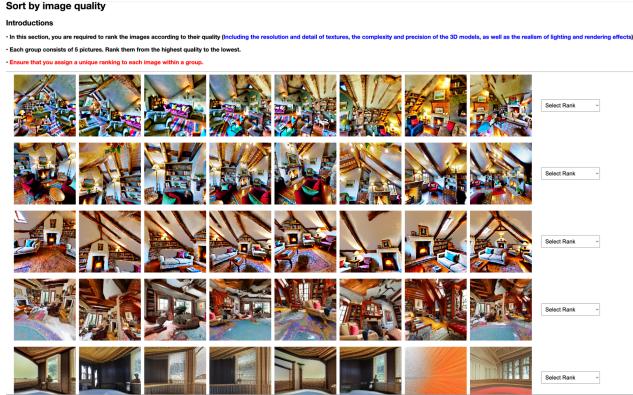Table 3. Time comparisons of ShowRoom3D and other state-of-the-art approaches.



Figure 9. User study interface.

dress oversaturation and enhance content diversity within a single image. However, this approach results in generating more crowded content in one view, exacerbating the Janus problem. The attempt to maintain consistency with the prompt in each image intensifies the overcrowded appearance of the room. Additionally, the increased diversity in every image can lead to more inconsistencies in style between views.

**Text2Room.** Text2Room shares the same disadvantages with the aforementioned baselines. Text2Room employs various strategies to fill the generated mesh and ensure its 'waterproof' quality, including random camera sampling and Poisson Reconstruction. However, this approach introduces more stretching and blurring artifacts in the rendered images.

**Look Outside The Room(LOTR).** LOTR is a novel view synthesis work to generate the next image conditioned on the previous images. However, it has certain disadvantages. Firstly, it struggles to generate images when the rotation degree varies too much, making it ineffective for panoramic image generation. Secondly, LOTR is constrained to generating images in specific directions, and it can not produce corresponding images if the camera is moved backward, up, or down.

## D. More Results of ShowRoom3D

In this section, we present an additional set of 16 views for a scene to illustrate that our room can be rendered at any position as shown in Figure 8. Furthermore, we provide more results with 8 views to demonstrate that our method is capable of generating diverse types of rooms as shown in Figure 10.

## E. User Study

We employ our method and other baselines to generate 13 room-scale scenes based on 13 prompts. To ensure fairness, we utilize Amazon MTurk to recruit 24 participants who

An animated room brimming with whimsical details, including a quirky bookshelf, a magical lamp, a playful computer, a desk having a mind of its own, a sofa with a welcoming smile, chairs dancing, art and paintings that tell whimsical tales a dining table with a feast for the eyes, and windows that frame a colorful cartoon world outside.



A living room filled with, furniture and some large mirrors, tables and chandelier.



An old traditional room with a cozy fireplace, an attic filled with treasures, a warm lamp, a well-stocked bookshelf, a vintage desk, a comfortable sofa, inviting chairs, captivating art and paintings, a serene yoga mat

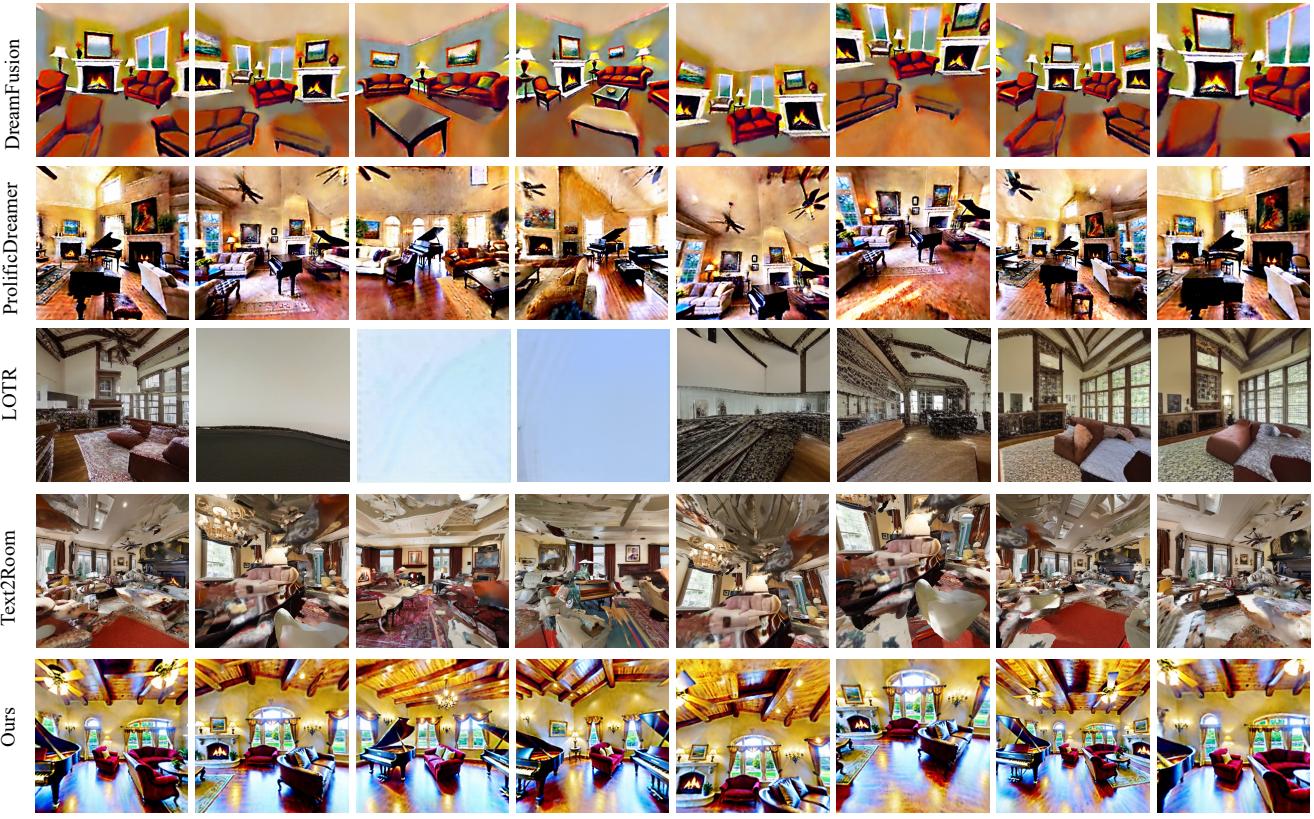| Azimuth: 0 Elevation: 0 | Azimuth : +90 Elevation : 0 | Azimuth : +180 Elevation : 0 | Azimuth : +270 Elevation : 0 | Azimuth : +90 Elevation : +20 | Azimuth : +90 Elevation : -20 | z = -1 | z = 1 |

Figure 10. Our more results with 8 views

rank the results on a scale from 5 (highest score) to 1 (lowest score), as depicted in Figure 9. Users are presented with multiple images from each scene.
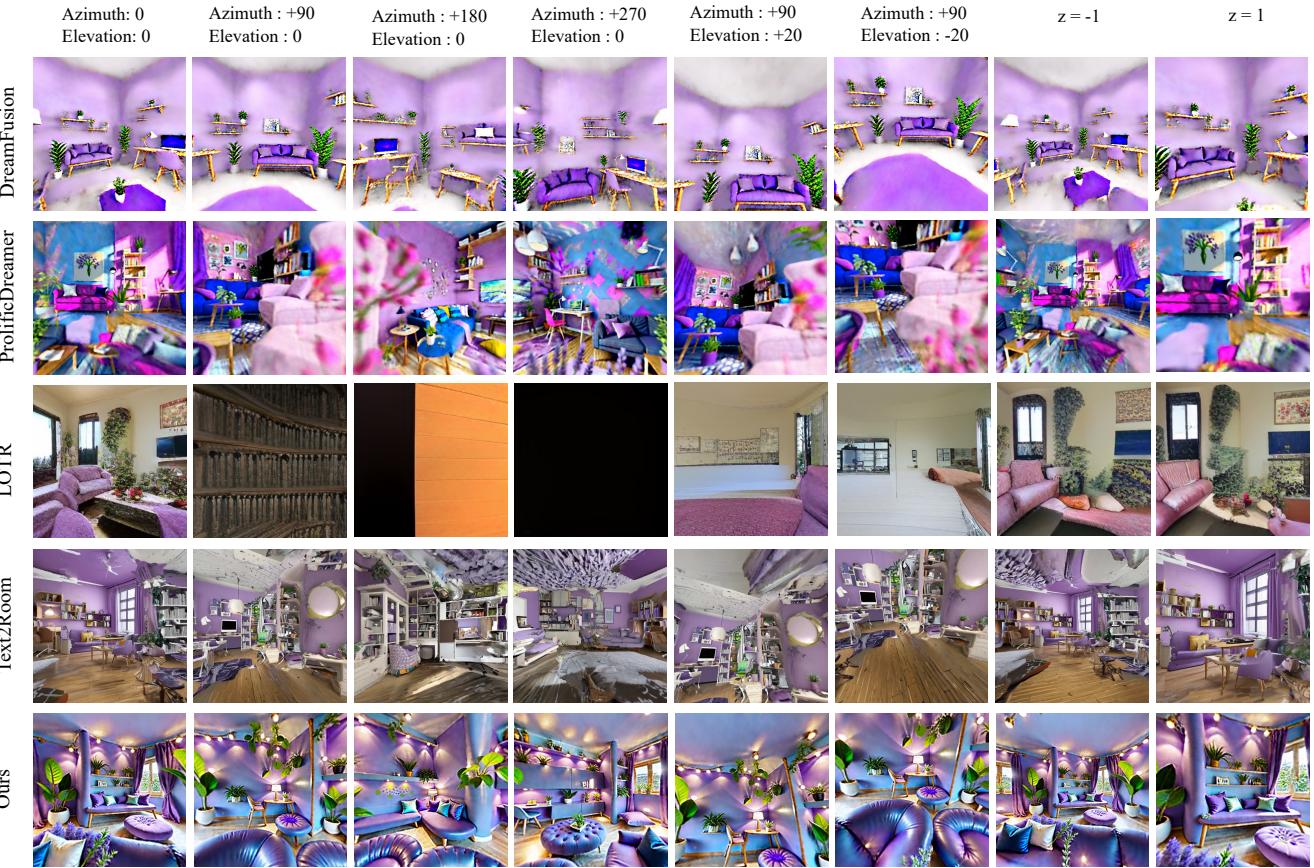
## F. Additional Ablation Study

We show another ablation study about SDS with our training tricks.

**Stable Diffusion In One Stage.** Now we show the results of Stable Diffusion (finetuned on Matterport3D dataset [3]) in one stage. The results will be shown in Figure 7. In Figure 7(a), we optimize the NeRF using the Stable Diffusion model (finetuned on the MatterPort3D dataset) in a single stage, incorporating all training tricks. Despite the inclusion of additional training tricks, proper room geometry generation remains elusive. This will demonstrate the effectiveness of our three-stage training pipeline in enhancing the geometry of the room, as shown in Figure 7(b).

A living room filled with furniture, grand pianos, fire places, paintings, large windows. A living room with couches and ceiling fans.

| Azimuth: 0<br>Elevation: 0 | Azimuth : +90<br>Elevation : 0 | Azimuth : +180<br>Elevation : 0 | Azimuth : +270<br>Elevation : 0 | Azimuth : +90<br>Elevation : +20 | Azimuth : +90<br>Elevation : -20 | z = -1 | z = 1 |

In lavender style, a versatile room with a sofa as the centerpiece, accompanied by various elements such as plants, a bookshelf, a lamp, a TV, games, a desk, a dining table, chairs, art and a meditation corner, providing a space for relaxation, reading, entertainment, work, gatherings, creativity, and tranquility.

| Azimuth: 0<br>Elevation: 0 | Azimuth : +90<br>Elevation : 0 | Azimuth : +180<br>Elevation : 0 | Azimuth : +270<br>Elevation : 0 | Azimuth : +90<br>Elevation : +20 | Azimuth : +90<br>Elevation : -20 | z = -1 | z = 1 |

Figure 11. Qualitative comparisons of ShowRoom3D and state-of-the-art approaches.