3D-SceneDreamer: Text-Driven 3D-Consistent Scene Generation

Songchun Zhang[1], Yibo Zhang[2], Quan Zheng[4], Rui Ma[2], Wei Hua[3]
Hujun Bao[1], Weiwei Xu[1], Changqing Zou[1,3*]

[1] Zhejiang University [2] Jilin University [3] Zhejiang Lab
[4] Institute of Software, Chinese Academy of Sciences

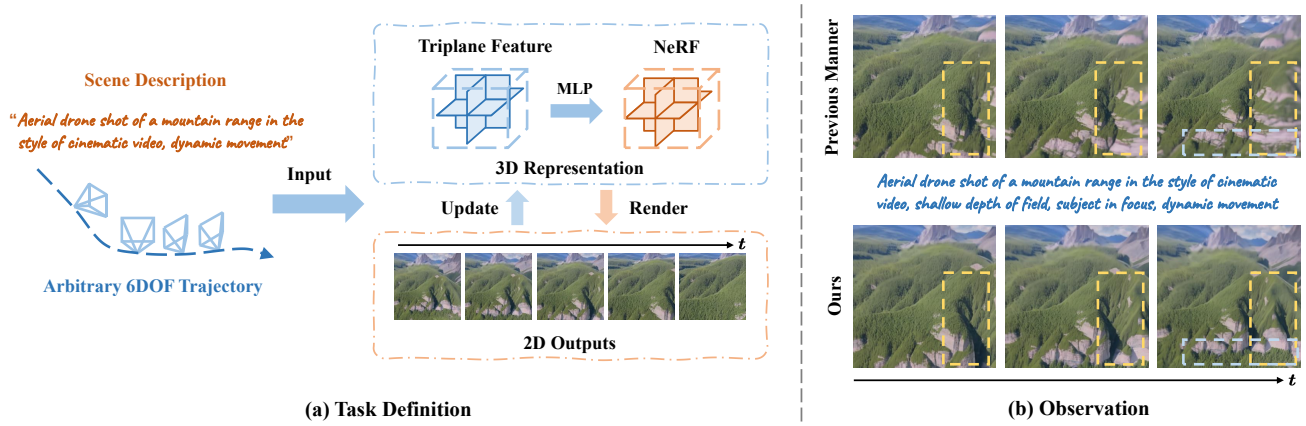


Figure 1. **Text-Driven 3D Scene Generation from text prompts.** (a) Given a scene description prompt and an arbitrary 6-degree-of-freedom (6-DOF) camera trajectory, our approach progressively generates the full 3D scene by continuously synthesizing 2D novel views. (b) The limitation of mesh representations [12, 16] and the lack of reasonable rectification mechanisms lead to cumulative errors in outdoor scenes, which are respectively marked with yellow and blue dash line boxes. In contrast, our approach can alleviate the problem by introducing a progressive generation pipeline.

Abstract

Text-driven 3D scene generation techniques have made rapid progress in recent years. Their success is mainly attributed to using existing generative models to iteratively perform image warping and inpainting to generate 3D scenes. However, these methods heavily rely on the outputs of existing models, leading to error accumulation in geometry and appearance that prevent the models from being used in various scenarios (e.g., outdoor and unreal scenarios). To address this limitation, we generatively refine the newly generated local views by querying and aggregating global 3D information, and then progressively generate the 3D scene. Specifically, we employ a tri-plane features-based NeRF as a unified representation of the 3D scene to constrain global 3D consistency, and propose a generative refinement network to synthesize new contents with higher quality by exploiting the natural image prior from 2D diffusion model as well as the global 3D information of the current scene. Our extensive experiments demonstrate that, in comparison to previous methods, our approach supports wide variety of scene generation and arbitrary camera trajectories with improved visual quality and 3D consistency.

1. Introduction

In recent years, with the growing need for 3D creation tools for metaverse applications, attention to 3D scene generation techniques has increased rapidly. Existing tools [11, 44] usually require professional modeling skills and extensive manual labor, which is time-consuming and inefficient. To facilitate the 3D scene creation and reduce the need for professional skills, 3D scene generation tools should be intuitive and versatile while ensuring sufficient controllability.

This paper focuses on the specific setting of generating consistent 3D scenes from the input texts that describe the 3D scenes. This problem is highly challenging from several perspectives, including the limitation of available text-3D data pairs and the need for ensuring both semantic and geometric consistency of the generated scenes. To overcome the limited 3D data issue, recent text-to-3D methods [42, 62] have leveraged the powerful pre-trained text-to-image diffusion model [48] as a strong prior to optimize 3D representation. However, their generated scenes often have relatively simpler geometry and lack 3D consistency, because 2D prior diffusion models lack the perception of 3D information.

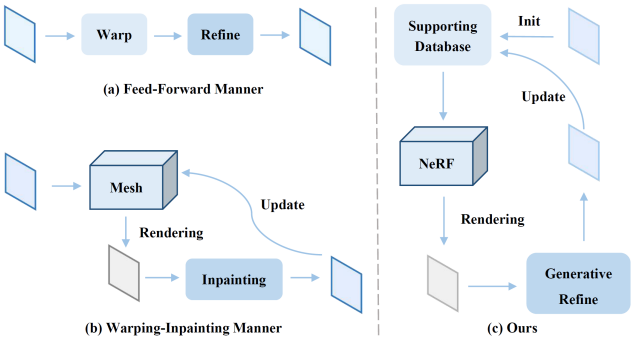


Figure 2. **Comparison with existing designs.** (a) The feed-forward approaches use depth-based warping and refinement operations to generate novel views of the scene without a unified representation. (b) The warping-inpainting approaches use mesh as a unified representation and generate the scene through iterative inpainting. (c) We replace the mesh with NeRF as the unified representation and alleviate the cumulative error issue by incorporating a generative refinement model. This allows our framework to support the generation of a wider range of scene types.

Some recent methods [12, 16] introduce the monocular depth estimation model [45, 46] as a strong geometric prior and follow the *warping-inpainting* pipeline [26, 29] for progressive 3D scene generation, which partially solves the inconsistency problem. Although these methods can generate realistic scenes with multi-view 3D consistency, they mainly focus on indoor scenes and fail to handle large-scale outdoor scene generation as illustrated in Fig. 1 (b). This can be attributed to two main aspects: (1) Due to the adoption of an explicit 3D mesh as the unified 3D representation, the noise of the depth estimation in the outdoor scene can cause a large stretch of the scene geometry; (2) The lack of an efficient rectification mechanism in the pipeline leads to an accumulation of geometric and appearance errors.

In this paper, we present a new framework, named **3D-SceneDreamer** that provides a unified solution for text-driven 3D consistent indoor and outdoor scene generation. Our approach employs a tri-planar feature-based radiance field as a unified 3D representation instead of 3D mesh, which is advantageous for general scene generation (especially in outdoor scenes) and supports navigating with arbitrary 6-DOF camera trajectories. Afterwards, we model the scene generation process as a progressive optimization of the NeRF representation, while a text-guided and scene-adapted generative novel view synthesis is employed to refine the NeRF optimization. Fig. 2 shows a comparison of our design with existing text-to-scene pipelines.

Specifically, we first perform scene initialization, which consists of two stages, i.e., generating a supporting database and optimizing the initial scene representation. We first use the input text prompt and the pre-trained diffusion model [48] to generate the initial image as an appearance prior. Then, we use an off-the-shelf depth estimation model [2] to provide the geometric prior for the corresponding scene. Inspired by [66], to prevent NeRF from over-fitting for the single view image, we construct a database via differentiable spatial transformation [18] and use it for optimizing the initial NeRF representation of the generated scene. To generate the extrapolated content, we use volume rendering and trilinear interpolation in the novel viewpoints to obtain the initial rendered images and their corresponding feature maps. These outputs are later fed into our 3D-aware generative refinement model, whose output images are subsequently added as new content to the supporting database. Next, in conjunction with the new data, we progressively generate the whole 3D scene by updating our 3D representation through our incremental training strategy.

Extensive experiments demonstrate that our approach significantly outperforms the state-of-the-art text-driven 3D scene generation method in both visual quality and 3D consistency. To summarize, our technical contributions are as follows:

- We provide a unified solution for text-driven consistent 3D scene generation that supports both indoor and outdoor scenes as well as allows navigation with arbitrary 6-DOF camera trajectories.
- We propose to use a tri-planar feature-based neural radiance field as a global 3D representation of the scene to generate continuous scene views, which preserves the 3D consistency of the scene, empowered by a progressive optimization strategy.
- We propose a new generative refinement model, which explicitly injects 3D information to refine the coarse view generated by novel view synthesis and then incorporates the new views to refine the NeRF optimization.

2. Related Work

Text-Driven 3D Content Generation. Recently, motivated by the success of text-to-image models, employing pre-trained 2D diffusion models to perform text-to-3D generation has gained significant research attention. Some pioneering works [42, 61] introduce the Score Distillation Sampling (SDS) and utilize 2D diffusion prior to optimize 3D representation. Subsequent works [8, 28, 34, 62] further enhance texture realism and geometric quality. However, they primarily focus on improving object-level 3D content generation rather than large-scale 3D scenes. Recent works [12, 16, 66] have proposed some feasible solutions for 3D scene generation. By utilizing the pre-trained monocular depth model and the inpainting model, they generate the 3D scene progressively based on the input text and camera trajectory. However, due to the underlying 3D representation or optimization scheme, these methods are limited in several aspects. For example, as [12, 16] utilize explicit mesh as 3D representation, it is difficult for them to generate outdoor scenes. Besides, their mesh outputs

also suffer from fragmented geometry and artifacts due to imprecise depth estimation results. Although Text2NeRF achieves to generate high-quality indoor and outdoor scenes by replacing the meshes with neural radiance fields [35], it can only generate camera-centric scenes. In contrast, our approach not only supports more general 3D scene generation but can also handle arbitrary 6DOF camera trajectories.

Text-Driven Video Generation. Text-Driven Video Generation aims to create realistic video content based on textual conditions. In the early stages, this task was approached using GAN [1, 25, 41] and VAE [33, 38] generative models, but the results were limited to low-resolution short video clips. Following the significant advancements in text-to-image models, recent text-to-video works extend text-to-image models such as transformer [17, 64, 65] and diffusion model [3, 14, 15, 32, 53, 68] for video generation. These approaches enable the generalization of high-quality and open-vocabulary videos, but require a substantial amount of text-image or text-video pairs of data for training. Text2Video-Zero [19] proposes the first zero-shot text-to-video generation pipeline that does not rely on training or optimization, but their generated videos lack smoothness and 3D consistency. Our method is capable of generating smooth and long videos wihch are consistent to the scenes described by the input text, without the need for large-scale training data. Furthermore, the utilization of NeRF as the 3D representation enhances the 3D consistency of our videos.

View Synthesis with Generative Models. Several early stage studies [5, 21, 22, 26, 29, 63] employ GAN to synthesize new viewpoints. However, the training process of GAN is prone to the issue of mode collapse, which limits the diversity of generation results. Diffusion model has been shown its capability to generate diverse and high-quality images and videos. In recent view synthesis works [4, 7, 51, 57], diffusion models have been employed to achieve improved scene generation results over prior works. For example, in Deceptive-NeRF [30], pseudo-observations are synthesized by diffusion models and these observations are further utilized for enhance the NeRF optimization. Closely similar to [30], our method propose a geometry-aware diffusion refinement model to reduce the artifacts of the input coarse view generated by the initial novel view synthesis. With the 3D information from NeRF features injected to the refinement process, we can achieve globally consistent 3D scene generation.

3. Neural Radiance Fields Revisited

Neural Radiance Fields (NeRF) [59] is a novel view synthesis technique that has shown impressive results. It represents the specific 3D scene via an implicit function, denoted as $f_\theta : (\boldsymbol{x}, \boldsymbol{d}) \mapsto (\mathbf{c}, \sigma)$, given a spatial location $\mathbf{x}$ and a ray direction $\mathbf{d}$, where θ represents the learnable parameters, and $\mathbf{c}$ and σ are the color and density. To render a novel image, NeRF marches a camera ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ starting from the origin $\mathbf{o}$ through each pixel and calculates its color $\hat{\boldsymbol{C}}$ and rendered depth $\hat{\boldsymbol{D}}$ via the volume rendering quadrature, *i.e.*, $\hat{\boldsymbol{C}}(\mathbf{r}) = \sum_{i=1}^{N} T_i \alpha_i \mathbf{c}_i$ and $\hat{\boldsymbol{D}}(\mathbf{r}) = \sum_{i=1}^{N} T_i \alpha_i t_i$, where $T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right)$, $\alpha_i = (1 - \exp(-\sigma_i \delta_i))$, and $\delta_k = t_{k+1} - t_k$ indicates the distance between two point samples. Typically, stratified sampling is used to select the point samples $\{t_i\}_{i=1}^{N}$ between t_n and t_f, which denote the near and far planes of the camera. When multi-view images are available, θ can be easily optimized with the MSE loss:

$$\mathcal{L}_\theta = \sum_{\boldsymbol{r} \in \mathcal{R}} \left\| \hat{\boldsymbol{C}}(\boldsymbol{r}) - \boldsymbol{C}(\boldsymbol{r}) \right\|_2^2 \tag{1}$$

where $\mathcal{R}$ is the collection of rays, and $\boldsymbol{C}$ indicates the ground truth color.

4. Methods

4.1. Overview

Given a description of the target scene a the input text prompt $\mathbf{p}$, and a pre-defined camera trajectory denoted by $\{\mathbf{T}_i\}_{i=1}^{N}$, our goal is to generate a 3D scene along the camera trajectory with the multiview 3D consistency.

The overview of the proposed model is illustrated in Fig. 3. We first introduce the acquisition of appearance and structural priors in Sec. 4.2, which serve as the scene initialization. The formulation of Unified Scene Representation and its optimization with the former priors are presented in Sec. 4.3. To synthesize new content while maintaining the multiview consistency, we propose a geometry-aware refinement model in Sec. 4.4. Finally, the full online scene generation process is presented in Sec. 4.5.

4.2. Scene Context Initialization

Given the input textual prompt $\mathbf{p}$, we first utilize a pre-trained stable diffusion model to generate an initial 2D image $\mathbf{I}_0$, which serves as an appearance prior for the scene. Then, we feed this image into the off-the-shelf depth estimation model [2], and take the output as a geometric prior for the target scene, denoted as $\mathbf{D}_0$. Inspired by [66], we construct a supporting database $\mathcal{S} = \{(\mathbf{D}_i, \mathbf{I}_i, \mathbf{T}_i)\}_{i=1}^{N}$ via differentiable spatial transformation [18] and image inpainting [16] techniques, where N denotes the number of initial viewpoints. This database provides additional views and depth information, which could prevent the model from overfitting to the initial view. With the initial supporting database, we can initialize the global 3D representation. The data generated by our method will be continuously appended to this supporting database for continuous optimization of the global 3D representation. More details are provided in our supplemental materials.

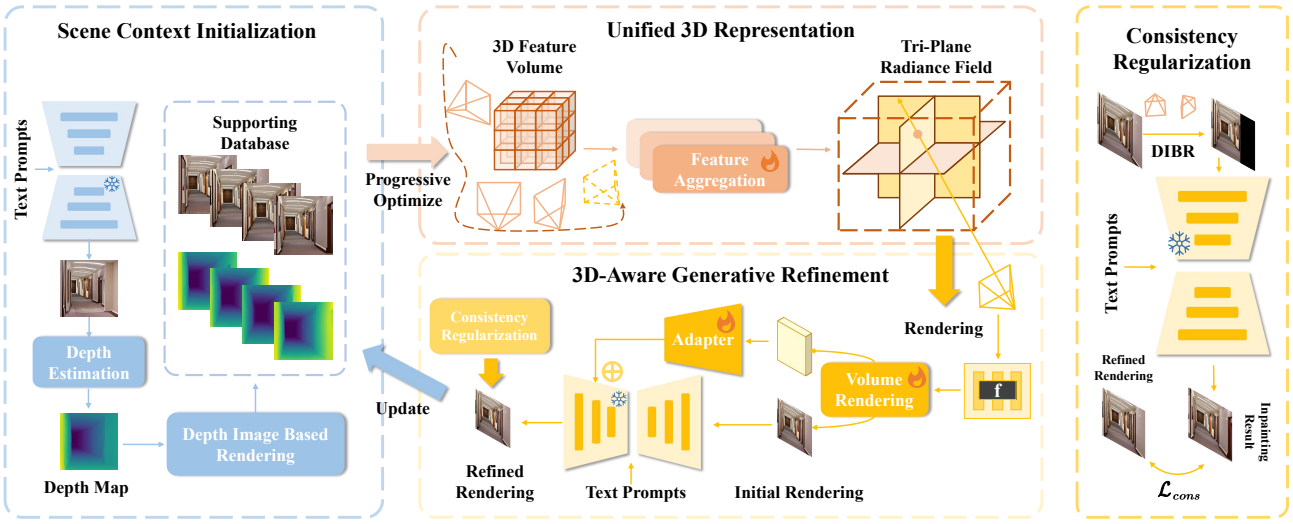


Figure 3. **Overview of our pipeline.** (a) **Scene Context Initialization** contains a supporting database to provide novel viewpoint data for progressive generation. (b) **Unified 3D Representation** provides a unified representation for the generated scene, which allows our approach to accomplish more general scene generation and to hold the 3D consistency at the same time. (c) **3D-Aware Generative Refinement** alleviates the cumulative error issue during long-term extrapolation by exploiting large-scale natural images prior to generatively refine the synthesized novel viewpoint image. The consistency regularization module is used for test-time optimization.

4.3. Unified Scene Representation

Though previous methods [26, 29] have achieved novel view generations via differentiable rendering-based frame-to-frame warping, there are still drawbacks: (1) the global 3D consistency is not ensured, (2) cumulative errors occur in long-term generation, (3) complex scenes may lead to failure. To tackling above issues, we propose a tri-planar feature-based NeRF as the unified representation. Compared with previous methods [12, 16, 26, 29], our approach constrains the global 3D consistency while handling the scene generation with complex appearances and geometries.

Tri-planar Feature Representation. For constructing the feature tri-planes $\mathbf{M} = \{\mathbf{M}_{xy}, \mathbf{M}_{yz}, \mathbf{M}_{xz}\} \in \mathbb{R}^{3\times S\times S\times D}$ from the input images, where S is the spatial resolution and D is the feature dimension, we first extract 2D image features from supporting views using the pre-trained ViT from DINOv2 [40] because of its strong capability in modeling cross-view correlations. We denote the extracted feature corresponding to image $\mathbf{I}_i$ as $\mathbf{F}_i$, and the feature set obtained from all input views is denoted as $\{\mathbf{F}_i\}_{i=1}^{N}$. To lift the local 2D feature maps into the unified 3D space, similar to the previous work [67], we back-project the extracted local image features $\mathbf{F}$ into a 3D feature volume $\mathbf{V}$ along each camera ray. To avoid the cubic computational complexity of volumes, we construct a tri-planar representation by projecting the 3D feature volume $\mathbf{V}$ into its respective plane via three separate encoders. This representation reduces the complexity from feature dimensionality reduction, but with equivalent information compared to purely 2D feature representations (e.g., BEV representations [10, 27]).

Implicit Radiance Field Decoder. Based on the constructed tri-planar representation $\mathbf{M}$, we can reconstruct the images with target poses via our implicit radiance field decoder module $\Psi = \{f_g, f_c\}$, where f_g and f_c indicate the geometric feature decoder and appearance decoder. Given a 3D point $p = [i, j, k]$ and a view direction $\boldsymbol{d}$, we orthogonally project p to each feature plane in $\mathbf{M}$ with bilinear sampling to obtain the conditional feature $\mathbf{M}_p = [\mathbf{M}_{xy}(i,j), \mathbf{M}_{yz}(j,k), \mathbf{M}_{xz}(i,k)]$. We feed $\mathbf{M}_p$ into the geometric feature decoder to obtain the predicted density σ and the geometric feature vector $\boldsymbol{g}$, after which we further decode its color $\boldsymbol{c}$:

$$\begin{aligned}(\sigma, \boldsymbol{g}) &= f_g\left(\gamma(\boldsymbol{x}), \mathbf{M}_p\right) \\ \boldsymbol{c} &= f_c\left(\gamma(\boldsymbol{x}), \gamma(\boldsymbol{d}), \boldsymbol{g}, \mathbf{M}_p\right)\end{aligned} \tag{2}$$

where $\gamma(\cdot)$ indicates the positional encoding function. Then we can calculate the pixel color via an approximation of the volume rendering integral mentioned in Sec. 3.

Training Objective. To optimize our 3D representation, we leverage the ground truth colors from the target image as the supervisory signal. Additionally, in the setting with sparse input views, we employ the estimated dense depth map to enhance the model's learning of low-frequency geometric information and prevent overfitting to appearance details. Our optimization objective is as follows:

$$\mathcal{L} = \sum_{\boldsymbol{r}\in\mathcal{R}} \left(\mathcal{L}_{photo}\left(\boldsymbol{r}\right) + \lambda \mathcal{L}_{depth}\left(\boldsymbol{r}\right)\right) \tag{3}$$

where $\mathcal{L}_{photo}\left(\boldsymbol{r}\right) = \left\|\hat{\boldsymbol{C}}\left(\boldsymbol{r}\right) - \boldsymbol{C}\left(\boldsymbol{r}\right)\right\|^2$, $\mathcal{L}_{depth}\left(\boldsymbol{r}\right) = \left\|\hat{\mathbf{D}}^*_{\mathbf{r}}\left(\boldsymbol{r}\right) - \mathbf{D}^*\left(\boldsymbol{r}\right)\right\|^2$, $\mathcal{R}$ denotes the collection of rays gen-

erated from the images in the supporting database, λ indicates the balance weight of the depth loss, and $\mathbf{D}^*(\boldsymbol{r})$ and $\hat{\mathbf{D}}^*_{\mathbf{r}}(\boldsymbol{r})$ denote the rendered depth and the depth obtained from the pre-trained depth estimation model. Since monocular depths are not scale- and shift-invariant, both depths are normalized per frame.

4.4. 3D-Aware Generative Refinement

Given a sequence of poses and an initial viewpoint, previous methods [12, 16, 66] usually generate novel views by the *warping-inpainting* pipeline. Though these methods have achieved promising results, they suffer from two issues: (1) The lack of rectification mechanisms in these methods can lead to error accumulation. (2) The lack of explicit 3D information during the inpainting process of these methods can lead to insufficient 3D consistency. Therefore, we propose a 3D-Aware Generative Refinement model to alleviate the above issues. On the one hand, we introduce an efficient refinement mechanism to reduce the cumulative error in the novel view generation. On the other hand, we explicitly inject 3D information during the process of generating novel views to enhance 3D consistency. We will describe the model design below.

Model Design. Given a novel viewpoint with camera pose $\mathbf{T}_i$, the tri-planar features $\mathbf{M}$, we can obtain the rendered image $\mathbf{I}_r$, rendered depth $\mathbf{D}_r$ and the corresponding 2D feature map $\mathbf{F}_r$ via the radiance field decoder module Ψ and volume rendering. For convenience, we model the whole process with a mapping operator $\mathcal{F}_{ren} : \{\mathbf{T}_i, \mathbf{M}\} \mapsto \{\mathbf{I}_r, \mathbf{F}_r, \mathbf{D}_r\}$. Note that the feature map is computed similarly to the color and depth, *i.e.*, by numerical quadrature, and can be formulated as

$$\mathbf{F}_r(\mathbf{r}) = \sum_{i=1}^{N} T_i \left(1 - \exp\left(-\sigma_i \delta_i\right)\right) \boldsymbol{g}_i \tag{4}$$

where $\boldsymbol{g}_i$ indicates the feature vector decoded by f_g, and N denotes the total number of point samples on the ray $\boldsymbol{r}$.

Although the quality of the rendered coarse results may not be very high, they can still provide reasonable guidance for the extrapolated view generation according to the current scene. Based on this assumption, we propose to take the rendered image and the feature map as conditional inputs to a pre-trained 2D stable diffusion model and generate a refined synthetic image $\hat{\mathbf{I}}_r$ via fine-tuning the model, which allows to leverage natural image priors derived from internet-scale data. The process can be formulated as:

$$\hat{\mathbf{I}}_r = \mathcal{F}_{gen}(\mathbf{I}_r, \tau(\mathbf{p}), \mathcal{G}(\mathbf{F}_r)) \tag{5}$$

where $\mathcal{F}_{gen}$ denotes our generative refinement model, $\tau(\mathbf{p})$ indicates the input text embedding, and $\mathcal{G}$ denotes the feature adapter for learning the mapping from external control information to the internal knowledge in LDM.

Scene-Adapted Diffusion Model Fine-Tuning. For the scene generation task, we propose to leverage the rich 2D priors in the pre-trained latent diffusion model instead of training a new model from scratch. Thus, we jointly train the feature adapter, the radiance field decoder, and the feature aggregation layer, while keeping the parameters of stable diffusion fixed. The objective of the fine-tuning process is shown below:

$$\mathcal{L}_{AD} = \mathbb{E}_{t,\epsilon\sim\mathcal{N}(0,I)}\left[\left\|\epsilon_\theta\left(\boldsymbol{z}_t, t, \tau(\mathbf{p}), \mathbf{F}_r, \mathbf{I}_r\right) - \epsilon\right\|_2^2\right] \tag{6}$$

With the rendered feature map $\mathbf{F}_r$ containing information about the appearance and geometry, we can control the pre-trained text-to-image diffusion model to generate images that are consistent with the content of generated images from previous viewpoints. In addition, our model inherits the high-quality image generation ability of the stable diffusion model, which ensures the plausibility of the generated views. The pre-trained prior and our effective conditional adaptation enable our model to have generalization ability in novel scenes.

Global-Local Consistency Regularization. In the online generation process, though our model can rectify the coarse rendering results, we do not explicitly constrain the 3D consistency across views when synthesizing novel views. Therefore, we design a regularization term $\mathcal{L}_{cons}$ for test-time optimization, which shares the same formula as Eq. (6) to guarantee the plausibility of the generated novel views. Specifically, we expect that 3D consistency exists between novel views obtained from geometric projection using local geometric information (i.e., monocular depth estimation) and novel views generated using global geometric information (i.e., global tri-planar 3D representation). Thus, we simultaneously generate novel views based on the previous warping-and-inpainting pipeline and use them as supervisory signals to further fine-tune the feature adapter.

4.5. Online Scene Generation Process.

In this section, we introduce our online 3D scene generation process, which consists of three parts: scene representation initialization, extrapolation content synthesis, and incremental training strategy.

Scene Representation Initialization. Given the input textual prompt, we first generate an initial 2D image using a pre-trained stable diffusion model, after which we construct a supporting database $\mathcal{S}$ via the method mentioned in Sec. 4.2. Then, by exploiting the data from the database, as well as the photometric loss (Eq. (3)), we can optimize the unified representation. To prevent the model from overfitting to high-frequency details, we allow the model to learn low-frequency geometric information better by utilizing the depth priors. [60].

Extrapolated Content Synthesis. To generate the extrapolated content, we proceed by retrieving the next pose, de-

noted as $\mathbf{T}_i$, from the pose sequence $\{\mathbf{T}_i\}_{i=1}^{N}$. We then employ volumetric rendering to obtain a coarse view of the current viewpoint and the corresponding feature map. These rendered outputs are used as conditional inputs to our generative refinement model $\mathcal{F}_{gen}$ for generating a refined view. Due to the presence of a generative refinement mechanism, our extrapolation method mitigates the effects of cumulative errors. The refined view from the model $\mathcal{F}_{gen}$ is subsequently added to the supporting database $\mathcal{S}$ as new content.

Incremental Training Strategy. After obtaining the new content, we then need to update the unified representation. However, fine-tuning only on the newly generated data can lead to catastrophic forgetting, whereas fine-tuning on the entire dataset requires excessively long training time. Inspired by [54], we sample a sparse set of rays $\mathcal{Q}$ according to the information gain to optimize the representation, thus improving the efficiency of the incremental training.

5. Experiments

5.1. Implementation details.

We implemented our system using PyTorch. For the differentiable rendering part, we utilized [13] for depth estimation. To avoid the occurrence of black holes, we referred to the implementation in [18] to generate surrounding views. For the text-guided image generation, we use the publicly available stable diffusion code from Diffusers [58]. For the multi-view consistency image generation, we refer to the implementation of T2I-Adapter [39] to inject the depth feature conditions. In the progressive NeRF reconstruction part, we refer to the tri-planar implementation in [6]. We conducted all experiments using 4 NVIDIA RTX A100 GPUs for training and inference. More details can be found in our supplementary material.

5.2. Evaluation metrics.

Image quality. We evaluate the quality of our generated images using CLIP Score (CS), Inception Score (IS), Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [36] and Natural Image Quality Evaluator (NQIE) [37]. The Inception Score is based on the diversity and predictability of the generated images. CLIP Score uses a pre-trained CLIP model [43] to measure the similarity between text and images. Note that existing visual quality metrics such as FID cannot be used since the scenes generated by text-to-3D approaches do not exhibit the same underlying data distribution.

Multiview Consistency. Given a sequence of rendered images, we evaluate the multi-view consistency of our generated scene using Camera Error (CE), Depth Error (DE), and flow-warping error (FE) metrics. Motivated by [10, 12], we use COLMAP [50], a reliable SfM technique, to compute the camera trajectory and the sparse 3D point cloud. CE is computed by comparing the difference between the predicted trajectory and the given trajectory, and DE is computed by comparing the difference between the sparse depth map obtained by COLMAP and the estimated depth map. In addition, to account for temporal consistency, we follow [23] and use RAFT [56] to compute FE.

Method	3D Representation	3D Consistency			Visual Quality			
		DE↓	CE↓	SfM rate↑	CS↑	BRISQUE↓	NIQE↓	IS↑
Inf-Zero [26]	-	-	1.189	0.38	-	**21.43**	5.85	2.34
3DP [52]	LDI&Mesh	0.42	0.965	0.47	-	29.95	5.84	1.75
PixelSynth [47]	Point Cloud	0.36	0.732	0.52	-	36.74	4.98	1.28
ProlificDreamer [62]	NeRF	-	-	-	23.41	27.97	6.75	1.21
Text2Room [16]	Mesh	0.24	0.426	0.63	28.15	28.37	5.46	2.19
Scenescape [12]	Mesh	0.18	0.394	0.76	28.84	24.54	4.78	2.23
Ours	NeRF	**0.13**	**0.176**	**0.89**	**29.97**	23.64	**4.66**	**2.62**

Table 1. **Comparison with text-to-scene methods.** We compare our approach with two categories of approaches, *i.e.*, pure text-driven 3D generation and text-to-image generation followed by 3D scene generation. Metrics on 3D consistency and visual quality are illustrated.

Method	FE↓	CS↑	BRISQUE↓	NIQE↓	IS↑
VideoFusion [32]	0.039	23.54	27.39	5.94	2.21
GEN-2 [49]	0.032	27.54	25.65	5.24	2.38
Ours	**0.028**	**29.95**	**23.53**	**4.70**	**2.69**

Table 2. **Comparision with text-to-video methods**. Metrics on flow warping error (FE) and visual quality are illustrated.

Method	CS↑	BRISQUE↓	NIQE↓	IS↑
Text2Light [20]	26.16	49.26	6.15	2.54
MVDiffusion [42]	27.25	31.54	5.47	2.76
Ours	**28.12**	**24.15**	**4.96**	**2.79**

Table 3. **Comparision with text-to-panorama methods.** We compare our method with recent text-driven 3D generation methods [9, 55]. Metrics on visual quality are illustrated.

Method	DE↓	CE↓	SfM rate↑	CS↑	BRISQUE↓	NIQE↓
Full Model	**0.13**	**0.176**	**0.89**	**29.97**	**26.18**	**6.54**
W/o UR	0.46	0.764	0.41	22.71	27.95	5.81
W/o GRM	0.59	0.981	0.46	22.12	29.64	5.75
W/o CR	0.19	0.254	0.78	28.14	27.16	6.12

Table 4. **Ablations.** For brevity, we use UR, GRM, CR to denote *Unified Representation*, *Generative Refinement Model* and *Consistency Regularization*, respectively.

5.3. Comparisons

Baselines. Since there are only a few baselines directly related to our approach, we also take into account some methods with similar capabilities and construct their variants for comparison. Specifically, the following three categories of methods are included:

- **Text-to-Scene.** There exist techniques [12, 16] that generate 3D meshes iteratively by employing warping and inpainting processes, allowing for direct comparisons with our proposed methods. Moreover, image-guided 3D generation methods [24, 26, 47] are also available, wherein initial images can be produced using a T2I model. Subse-

Figure 4. **Quantitative Results.** From our results, it can be seen that our approach produces high-fidelity scenes with stable 3D consistency in indoor scenes, outdoor scenes, and unreal-style scenes. More high-resolution results can be found in the supplementary material.

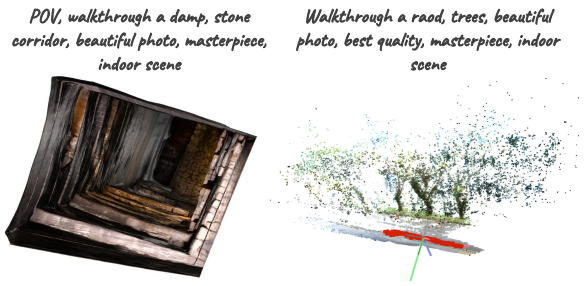
Figure 5. **Comparison with text-to-panorama methods**. It can be seen that although our method is not trained on panoramic data, it can also generate multiple views with cross-view consistency.

Figure 6. **Comparison with text-to-video methods.** Blur artifacts and temporally inconsistent frames occur in the text-to-video methods because of the lack of global 3D representation.

quently, their pipeline can be used to generate 3D scenes, enabling a comparison against our approach. We comprehensively evaluate these methods based on the previously introduced 3D consistency and visual quality metrics.

- **Text-to-Video.** Some recent text-driven video generation methods [32, 49] can also generate similar 3D scene walkthrough videos. Since it is not supported to explicitly control the camera motion in the video generation methods, we only evaluated them in terms of visual quality and temporal consistency.
- **Text-to-Panorama.** This task generates perspective images covering the panoramic field of view, which is challenging to ensure consistency in the overlapping regions. We have selected two related methods [9, 55] for comparisons.

Comparison to Text-to-Scene Methods. To generate the scenes, we use a set of test-specific prompts covering descriptions of indoor, outdoor and unreal scenes. Each prompt generates an image sequence of 100 frames, and for a fair comparison, we set a fixed random seed. After that, we compute the metrics proposed in Sec. 5.2 on the generated image sequences and evaluate the effectiveness of the method. As shown in Tab. 1, our method outperforms the mesh-based iterative generation methods in several metrics, especially for outdoor scenes. The quality of



(a) Extracted Mesh **(b) Point Cloud**

Figure 7. **Reconstructed 3D Results.** (a) The 3D mesh extracted by marching cube algorithm, and (b) the point cloud obtained after the reconstruction using COLMAP [31]. Our reconstruction results show that our methods can generate scenes with satisfactory 3D consistency.

their generation results relies heavily on the generative and geometric prior and degrades over time due to error accumulation. In addition, their use of a mesh to represent the scene makes it difficult to represent intense depth discontinuities, which are common in outdoor scenes. Our method, on the other hand, adopts hybrid NeRF as the scene representation, which can cope with complex scenes, and our rectification mechanism can mitigate the effect of accumulated errors caused by inaccurate prior signals.

Comparison to Text-to-Video Methods. For comparison with the text-to-video model, we used the same collection of prompts as input to the model and generated 1,200 video clips. We used the same metrics to evaluate the 3D consistency and visual quality of the videos generated by the T2V model and our rendered videos. As shown in Tab. 2, our method significantly outperforms the T2V model on all metrics, proving the effectiveness of our method. The T2V model learns geometry and appearance prior by training on a large video dataset, but it lacks a unified 3D representation, making it difficult to ensure multi-view consistency of the generated content, as can be observed Fig. 6.

Comparison to Text-to-Panorama Methods. We evaluate the methods [9, 55] on visual quality. Tab. 3 and Fig. 5 present the quantitative and qualitative evaluations, respectively. From the results, it can be seen that the results of previous methods can be inconsistent at the left and right boundaries, while our method, although not specifically designed for panorama generation, produces multiple views with cross-view consistency.

3D Results. In Fig. 7, we show the 3D results reconstructed by our method. The 3D mesh is extracted by the marching cube algorithm [31]. Additionally, we can reconstruct high-quality point clouds using colmap [50] by inputting the rendered image collection, which further demonstrates the superior 3D consistency of the generated view results.

5.4. Ablation Study

To further analyze the proposed methodology, we performed several ablation studies to evaluate the effectiveness of each module. More ablation studies can be found in our supplementary material.

Effectiveness of Unified Representations. To validate our necessity to construct a unified 3D representation, we remove it from our pipeline. At this time, our approach degenerates to the previous paradigm of warping-inpainting. As shown in Tab. 4, the quality of the generated scenes degrades in DE and CE metrics due to the lack of global 3D consistency constraints.

Effectiveness of Generative Refinement. To validate the effectiveness of our proposed generative refinement, we ablate the modules in our approach, whereby the novel view obtained through volume rendering will be updated directly into the supporting database for subsequent incremental training. The results in Tab. 4 show that this can lead to a significant degradation in the quality of the generated scene. We argue that the reason for this is that the quality of novel views generated by NeRF training on sparse views tends to be inferior, with notable blurring and artifacts. Therefore, adding this data for optimizing 3D scenes would lead to continuous degradation of the quality of the generated scenes.

Effectiveness of Consistency Regularization. To verify the validity of our regularization loss, we ablate this loss and generate scenes to compute the relevant metrics. As shown in Tab. 4, adding this loss further improves the 3D consistency of the generated scenes. Though we explicitly inject 3D information into the refining process, its output still shows some inconsistent results in several scenes. Therefore, to further improve the quality of the generated new views, we perform test-time optimization through this regularization term to constrain the consistency between local and global representations.

6. Conclusion

This paper presents a new framework, which employs the tri-planar feature-based neural radiation field as a unified 3D representation and provides a unified solution for text-driven indoor and outdoor scene generation and the output supports navigation with arbitrary camera trajectories. Our method fine-tunes a scene-adapted diffusion model to correct the generated new content to mitigate the effect of cumulative errors while synthesizing extrapolated content. Experimental results show that our method can produce results with better visual quality and 3D consistency compared to previous methods.

7. Acknowledgements

This research was supported by Zhejiang Provincial Natural Science Foundation of China under Grant No. LD24F020007, National Natural Science Foundation of China (No. 62202199), and NSFC (no. 62302491).

References

[1] Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. Conditional GAN with discriminative filter generation for text-to-video synthesis. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1995–2001, 2019. 3

[2] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 2, 3

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 3

[4] Shengqu Cai, Eric Ryan Chan, Songyou Peng, Mohamad Shahbazi, Anton Obukhov, Luc Van Gool, and Gordon Wetzstein. Diffdreamer: Consistent single-view perpetual view generation with conditional diffusion models. *arXiv preprint arXiv:2211.12131*, 2022. 3

[5] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5799–5809, 2021. 3

[6] 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, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 6

[7] Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. *arXiv preprint arXiv:2304.02602*, 2023. 3

[8] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2

[9] Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. Text2light: Zero-shot text-driven hdr panorama generation. *ACM Transactions on Graphics (TOG)*, 41(6):1–16, 2022. 6, 7, 8

[10] Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. Scenedreamer: Unbounded 3d scene generation from 2d image collections. *arXiv preprint arXiv:2302.01330*, 2023. 4, 6

[11] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 1

[12] Rafail Fridman, Amit Abecasis, Yoni Kasten, and Tali Dekel. Scenescape: Text-driven consistent scene generation. *arXiv preprint arXiv:2302.01133*, 2023. 1, 2, 4, 5, 6

[13] Vitor Guizilini, Igor Vasiljevic, Dian Chen, Rareș Ambruș, and Adrien Gaidon. Towards zero-shot scale-aware monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9233–9243, 2023. 6

[14] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 3

[15] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 3

[16] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7909–7920, 2023. 1, 2, 3, 4, 5, 6

[17] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 3

[18] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015. 2, 3, 6

[19] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023. 3

[20] Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. CLIP-mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia 2022 Conference Papers*. ACM, 2022. 6

[21] Jing Yu Koh, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. Pathdreamer: A world model for indoor navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14738–14748, 2021. 3

[22] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1169–1178, 2023. 3

[23] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 170–185, 2018. 6

[24] Xingyi Li, Zhiguo Cao, Huiqiang Sun, Jianming Zhang, Ke Xian, and Guosheng Lin. 3d cinemagraphy from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4595–4605, 2023. 6

[25] Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 3

[26] Zhengqi Li, Qianqian Wang, Noah Snavely, and Angjoo Kanazawa. Infinitenature-zero: Learning perpetual view generation of natural scenes from single images. In *Proceed-*

ings of the European Conference on Computer Vision, pages 515–534. Springer, 2022. 2, 3, 4, 6

[27] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *Proceedings of the European Conference on Computer Vision*, pages 1–18. Springer, 2022. 4

[28] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023. 2

[29] Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14458–14467, 2021. 2, 3, 4

[30] Xinhang Liu, Shiu-hong Kao, Jiaben Chen, Yu-Wing Tai, and Chi-Keung Tang. Deceptive-nerf: Enhancing nerf reconstruction using pseudo-observations from diffusion models. *arXiv preprint arXiv:2305.15171*, 2023. 3

[31] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pages 347–353. 1998. 8

[32] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10209–10218, 2023. 3, 6, 7

[33] Tanya Marwah, Gaurav Mittal, and Vineeth N Balasubramanian. Attentive semantic video generation using captions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1426–1434, 2017. 3

[34] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. *arXiv preprint arXiv:2211.07600*, 2022. 2

[35] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3

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

[37] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 6

[38] Gaurav Mittal, Tanya Marwah, and Vineeth N Balasubramanian. Sync-draw: Automatic video generation using deep recurrent attentive architectures. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1096–1104, 2017. 3

[39] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 6

[40] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4

[41] Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. To create what you tell: Generating videos from captions. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1789–1798, 2017. 3

[42] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022. 1, 2, 6

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6

[44] Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, Alejandro Newell, Hei Law, Ankit Goyal, Kaiyu Yang, and Jia Deng. Infinite photorealistic worlds using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12630–12641. IEEE, 2023. 1

[45] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 2

[46] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 2

[47] Chris Rockwell, David F Fouhey, and Justin Johnson. Pixelsynth: Generating a 3d-consistent experience from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14104–14113, 2021. 6

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2

[49] RunWay. Gen-2: The next step forward for generative ai, 2023. https://research.runwayml.com/gen2. 6, 7

[50] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 6, 8

[51] Liao Shen, Xingyi Li, Huiqiang Sun, Juewen Peng, Ke Xian, Zhiguo Cao, and Guosheng Lin. Make-it-4d: Synthesizing

a consistent long-term dynamic scene video from a single image. *arXiv preprint arXiv:2308.10257*, 2023. 3

[52] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8028–8038, 2020. 6

[53] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 3

[54] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J Davison. imap: Implicit mapping and positioning in real-time. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6229–6238, 2021. 6

[55] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *arXiv preprint arXiv:2307.01097*, 2023. 6, 7, 8

[56] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision*, pages 402–419. Springer, 2020. 6

[57] Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16773–16783, 2023. 3

[58] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 6

[59] Can Wang, Ruixiang Jiang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Nerf-art: Text-driven neural radiance fields stylization. *arXiv preprint arXiv:2212.08070*, 2022. 3

[60] Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. *arXiv preprint arXiv:2303.16196*, 2023. 5

[61] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023. 2

[62] 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, 2023. 1, 2, 6

[63] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7467–7477, 2020. 3

[64] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021. 3

[65] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. In *Proceedings of the European Conference on Computer Vision*, pages 720–736. Springer, 2022. 3

[66] Jingbo Zhang, Xiaoyu Li, Ziyu Wan, Can Wang, and Jing Liao. Text2nerf: Text-driven 3d scene generation with neural radiance fields. *arXiv preprint arXiv:2305.11588*, 2023. 2, 3, 5

[67] Xiaoshuai Zhang, Sai Bi, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. Nerfusion: Fusing radiance fields for large-scale scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5449–5458, 2022. 4

[68] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 3