

CompoNeRF: Text-guided Multi-object Compositional NeRF with Editable 3D Scene Layout

Yiqi Lin^{1*} Haotian Bai^{1*} Sijia Li² Haonan Lu² Xiaodong Lin³ Hui Xiong^{1,4} Lin Wang^{1,4} †
¹AI Thrust, HKUST(Guangzhou) ²OPPO ³Rutgers University ⁴Dept. of CSE, HKUST
ylin933@connect.hkust-gz.edu.cn haotianwhite@outlook.com {lisijia, luhaonan}@oppo.com
lin@business.rutgers.edu {xionghui, linwang}@ust.hk

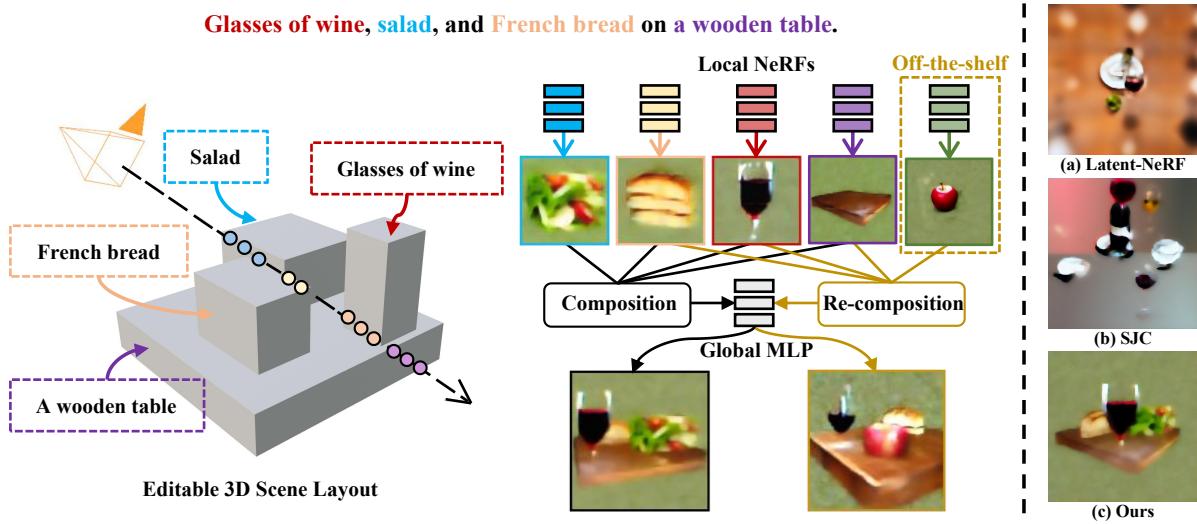


Figure 1: **Left:** CompoNeRF generates multi-object 3D scenes based on the editable 3D scene layout that represents each object with a local NeRF associated with its spatial location (*i.e.*, 3D box) and local text prompt. Also, the 3D scene layout makes the object-compositional scene generation flexibly editable (*e.g.*, scaling, moving, duplication, or re-composition) by manipulating the 3D layout or text prompt. **Right:** Our approach produces the most faithful and editable scenes given the multi-object text guidance, while the Latent-NeRF [19] and SJC [41] suffer from missing objects and semantic confusion.

Abstract

Recent research endeavors have shown that combining neural radiance fields (NeRFs) with pre-trained diffusion models holds great potential for text-to-3D generation. However, a hurdle is that they often encounter guidance collapse when rendering complex scenes from multi-object texts. Because the text-to-image diffusion models are inherently unconstrained, making them less competent to accurately associate object semantics with specific 3D structures. To address this issue, we propose a novel framework, dubbed **CompoNeRF**, that explicitly incorporates an editable 3D scene layout to provide effective guidance at the

single object (*i.e.*, local) and whole scene (*i.e.*, global) levels. Firstly, we interpret the multi-object text as an editable 3D scene layout containing multiple local NeRFs associated with the object-specific 3D box coordinates and text prompt, which can be easily collected from users. Then, we introduce a global MLP to calibrate the compositional latent features from local NeRFs, which surprisingly improves the view consistency across different local NeRFs. Lastly, we apply the text guidance on global and local levels through their corresponding views to avoid guidance ambiguity. This way, our CompoNeRF allows for flexible scene editing and re-composition of trained local NeRFs into a new scene by manipulating the 3D layout or text prompt. Leveraging the open-source Stable Diffusion model, our CompoNeRF can generate faithful and editable text-to-3D

*Equal contribution † Corresponding author

results while opening a potential direction for text-guided multi-object composition via the editable 3D scene layout.

1. Introduction

Recently, text-to-image generation [9, 25, 31] has achieved tremendous success by coupling the vision-language pre-trained models [30, 14] with diffusion models [9, 25, 31]. These breakthroughs have also yielded far-reaching implications in text-to-3D generation [11, 34, 10, 7, 23, 13, 45] using powerful vision-language pre-trained models. More recently, several text-to-3D methods [28, 15, 19, 41] have shown that matching the rendered views from the differential 3D model, such as Neural Radiance Fields (NeRFs) [20, 3, 24], with the learned text-to-image distribution from pre-trained diffusion model can achieve remarkable results.

However, the textual description is often an abstract specification for a desired target 3D model or a 2D image. Despite that the powerful diffusion models, *e.g.*, Stable Diffusion [19], have been trained on billions of text-image pairs [35], it is still a challenge to generate geometrically coherent images across different viewpoints from the text. Moreover, the diffusion model may produce inaccurate results [5] given text containing multiple objects, resulting in missing objects or semantic confusion. For example, Fig. 2 demonstrates that the Stable Diffusion fails to maintain the object identities and geometric coherence even with the simple multi-object text. This obviously contradicts the essence of volume rendering in NeRF, leading to a hurdle —*guidance collapse*, especially when rendering complex scenes from multi-object texts. As a result, the state-of-the-art (SoTA) Latent-NeRF [19] and SJC [41] models can only generate part of concepts in the multi-object text, as shown in Fig. 2, limiting its application towards object-compositional 3D scene generation from the text prompt.

Therefore, it naturally raises the question: *whether all the concepts in the multi-object text can be accurately learned and composed from the agnostic distribution of the diffusion model for 3D scene generation*. As shown in Fig. 2, we observe that the diffusion model can more accurately generate single objects with their respective local text prompts. This motivates us to introduce more fine-grained text guidance to tackle the guidance collapse issue in existing frameworks [19, 41] when taking multi-object text prompts. Thus, a straightforward solution is binding object-oriented guidance for each object into the particular 3D locations, making the 3D representation and rendering pipeline object-aware. However, existing approaches [28, 15, 19, 41] tend to encode the entire scene into a single neural network, making it hard to incorporate the decomposed guidance during training as they are generally agnostic to the object’s identity.

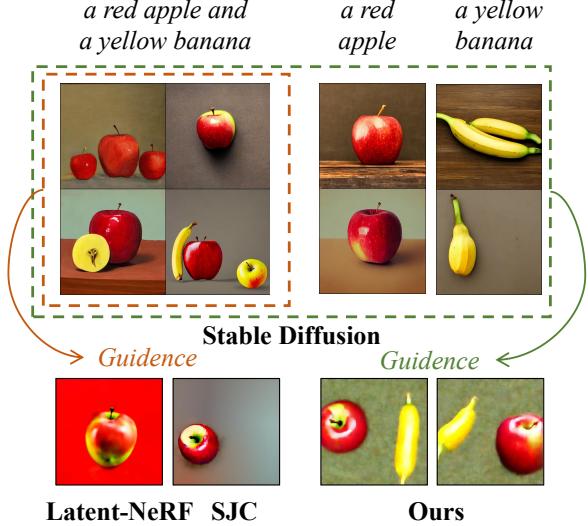


Figure 2: The guidance collapse issue on generating the multi-object scene associated with Stable Diffusion [31]. Comparison of our CompoNeRF with Latent-NeRF [22] and SJC [41] shows that using global (scene) and local (object) text guidance can mitigate the such problem.

To overcome these challenges, we propose a compositional NeRF framework, called **CompoNeRF**, in which the multi-object text guidance is interpreted as an editable 3D scene layout. As shown in Fig. 1, the scene layout collects individual object entities from the input text and gathers their pre-defined coarse 3D bounding box information, *e.g.*, box coordinates. In CompoNeRF, each box in the 3D scene layout is modeled by a local NeRF for representation learning, and the complete scene (*i.e.*, global) views are rendered by compositing all the learned 3D representations from local NeRFs. Nevertheless, the direct composition from all local NeRFs may not ensure learning coherent global views without tackling the two following issues.

Firstly, the global view consistency is hard to capture across multiple local NeRFs as the none shared parameters. Therefore, we use a global MLP to calibrate the local NeRF predictions by leveraging samples’ global coordinates and ray’s directions as conditional input. The model can gradually learn global consistency across multiple objects by passing the global information through the same global MLP to local NeRFs. Secondly, certain objects in the scene may be fully occluded in the training data due to random camera positions, leading to inaccurate text guidance. To tackle the occlusion issue, for each local NeRF, we separately apply local text guidance on the locally rendered view, as the diffusion model can provide more accurate guidance to shape the object identity despite occlusions. As a result, our approach ensures coherent and realistic 3D generation, even in crowded scenes, as demonstrated in Fig. 1, without missing objects or ambiguity. Moreover, the 3D scene layout facilitates object editing by allowing

Methods	Diffusion Model	3D Representation	Scene Rendering	Input Prompt	Scene Editing	Re-composition
DreamFusion [28]	Imagen [33]	Mip-NeRF 360 [4]	Object-centric	Text	T	✗
Magic3D [15]	eDiff-I [2] + SD [31]	Instant-NGP [24]	Object-centric	Text	T	✗
SJC [31]	SD [31]	voxel radiance field	Object-centric	Text	T	✗
Latent-NeRF [19]	SD [31]	Instant-NGP [24]	Object-centric	Text+Fine Shape	T	✗
Ours	SD [31]	Instant-NGP [24]	Object-compositional	Text+3D Coarse Boxes	T/M/S/R	✓

Table 1: Comparison of our method and the related works for text-to-image generation. SD denotes Stable Diffusion. For scene editing, we use T(editing object with text), M(moving object), S(scaling object), and R(remove object) for short.

users to modify text and manipulate objects flexibly, including moving, scaling, and duplicating. Given a vast pre-trained content gallery, users can rapidly generate their desired 3D scenes using text prompts and 3D scene layouts, thereby democratizing 3D content creation. The comparisons of editing approaches are summarized in Tab. 1.

To summarize, our paper makes three key contributions: **(I)** We address the guidance collapse issue in multi-object 3D scene generation by integrating an editable 3D layout with multiple local NeRFs to precisely associate guidance for specific structures. **(II)** We tackle the global consistency and occlusion issue by introducing a global MLP to calibrate the global scene color and different levels of text guidance to maintain the identity of objects while learning the global coherence for individual entities. **(III)** We thoroughly evaluate the effectiveness of our proposed method across various multi-object scenarios, demonstrating its ability to generate 3D scenes compositionally and offer flexible editing capabilities.

2. Related Work

Text-guided 3D Generative Models. Motivated by the success of the vision-language models, e.g., CLIP [30], text-guided 3D generation has also been rapidly progressing. Several works [11, 40, 23, 34, 45] utilize the pre-trained vision-language model to provide robust alignment between image features extracted from rendered views of 3D representations and text features by minimizing feature similarity. Though pre-trained large-scale vision-language models can offer strong 2D guidance for 3D representation learning, the quality of their 2D rendering results may be less realistic due to insufficient details in the features. Recently, DreamFusion [28] demonstrates remarkable capability in text-to-3D object generation by incorporating a powerful pre-trained text-to-image diffusion model [32] through score distillation sampling. Similarly, Magic3D [15] proposes a two-stage super-resolution method to enhance generation quality with hybrid 3D representation. Latent-NeRF [19] demonstrates that updating the NeRF in the latent space using score distillation sampling also can produce realistic results. SJC [41] improves the sampling process by introducing a perturb-and-average scoring scheme to address distribution mismatching issues.

By contrast, our CompoNeRF generates multi-object 3D scenes in an *object-compositional* way by utilizing multi-

ple local NeRFs to model the scene instead of a holistic object-centric representation. Our editable 3D scene layout allows for editing of scenes by changing the text prompt similarly as [28, 15]. Also, it enables manipulation of the spatial arrangement of individual objects in a crowded scene, as summarized in Tab. 1. Moreover, the layout allows for re-composition with other off-the-shelf representations, enabling the rapid generation of new scenes.

Neural Rendering for 3D Modeling. The rapid advancement of NeRF has greatly improved the performance of neural renderers. NeRF models [21, 16, 24, 17, 3, 39, 48] are a family of volume rendering algorithms that utilizes coordinate-based MLPs to directly predict color and opacity from the 3D position and 2D viewing direction. The photo-realistic synthesized views from these models have led to the widespread adoption of differential volume rendering in various applications, such as relighting [37, 49], dynamic scene reconstruction [6, 29, 44, 38], editable scenes and avatars [18, 47], and surface reconstruction [1, 42]. Most NeRF applications commonly use *single* multi-layer perceptron to encode the entire scene into parameters, which can be ambiguous w.r.t. the specific object identities within the scene. Note that our method involves rendering the scene using multiple local NeRFs, and the rendering process requires compositing all the local NeRF predictions based on their spatial relationships.

Object-Compositional Scene Modeling. The idea of generating new scenes by compositing multiple object-centric representations is a straightforward approach in scene generation [8]. Several works [50, 47, 43, 22, 26, 27, 46, 36] attempt to directly decompose object representations from the scene image to perform compositional scene modeling. Based on the additional object-level information, they can be roughly categorized as semantic based [50, 47, 43, 22, 26] and 3D layout based [27, 46, 36]. Semantic based methods incorporate extra semantic information, such as segmentation labels [50], object instance mask [47, 43], and features from the pre-trained vision-language model [22], to learn the object representations. On the other hand, 3D layout based methods directly use the 3D object coordinate information as the object guidance for learning object-specific representation and generating the full scene by compositing all object representations. For instance, NSG [27] and its variant [36] use a scene graph structure to model dynamic scenes, associating each object with a 3D box and a

node representation. While previous works focused on decomposing object entities from real-world images, our work serves as the *first* attempt to tackle the text-to-3D generation via decomposed representations with 3D scene layout.

3. Methodology

3.1. Overview

Fig. 3 illustrates our pipeline, which consists of three main components, including the editable 3D scene layout based on multi-object text (Sec. 3.3), the scene rendering pipeline that composites the predictions from all local NeRFs (Sec. 3.4), and the joint optimization on both local and global representation models (Sec. 3.5). To elaborate, our editable 3D scene layout represents a global frame of the scene by decomposing it into a set of local frames, where each is parameterized by a local NeRF, a 3D bounding box, and a corresponding local text prompt. For instance, the text prompt ‘A teddy bear and a stuffed monkey sit side by side’ is interpreted as a 3D scene layout, as shown in Fig. 3. The whole 3D layout, *i.e.*, scene frame, consists of two 3D bounding boxes, *i.e.* local frames #1 and #2, with specific local text prompts, *i.e.*, ‘a teddy bear’ and ‘a stuffed monkey’. To render the scene view, we first calculate the ray-box intersections between the boxes and rays $(\mathbf{r}_o, \phi_d, \theta_d)$, where the \mathbf{r}_o is the ray origin and the (\mathbf{r}_o, ϕ_d) is its direction. Then, to infer each object’s properties in local NeRFs, we sample the global points (x_g, y_g, z_g) in the global frame within the ray-box intersection intervals and project them into the normalized local location (x_l, y_l, z_l) in the local frame. Given the local sampling points (x_l, y_l, z_l) , the implicit local NeRF θ_l outputs four pseudo-color channels C_l and density σ , which can be used to render a local view of the local frame to match its local text prompt. We further calibrate the predicted pseudo-color C_l from local frames by adding the global embeddings emb_g to improve the global view consistency. Then, the calibrated predictions after composition are used to reconstruct the scene view by volumetric rendering along the rays. Lastly, the rendered views based on local and global frames are guided by score distillation sampling loss $\nabla \mathcal{L}_{\text{SDS}}$ [28] to optimize all the learnable parameters.

3.2. Preliminaries

3D Representation in Latent Space. Our approach is built upon the SoTA text-to-image model—Stable Diffusion [31]. To avoid heavy computation in the pixel space, we follow the Latent-NeRF [19] to model each object with an independent local NeRF θ_l that outputs four pseudo-color channels C , corresponding to the four latent features that Stable Diffusion operates over, and a volume density σ . Specifically, the representation maps a point $(x_l, y_l, z_l) \in [-1, 1]$ in the local frame to its corresponding volumetric density σ and emitted color C_l to the latent features, *i.e.*, $[C_l, \sigma] = \theta_l(x_l, y_l, z_l)$. The predicted pseudo-color is fed

forward into the decoder of Stable Diffusion models to obtain the final rendering result.

Score Distillation Sampling. To achieve text-to-3D generation, DreamFusion [28] introduces Score Distillation Sampling (SDS) to propagate the text-to-image generative prior from diffusion model ϕ to the NeRF parameters θ . During the SDS process, a noise image \mathbf{X}_t is first generated by adding a sampled noise $\epsilon \sim \mathcal{N}(0, I)$ in noise level t into a rendered view \mathbf{X} from a NeRF. Then, the diffusion model ϕ predicts the sampled noise $\epsilon_\phi(\mathbf{X}_t, t, T)$ given the noisy image \mathbf{X}_t , noise level t , and optional text prompt T . Specifically, SDS computes the gradient from the difference between the predicted and added noises,

$$\nabla_{\theta} \mathcal{L}_{\text{SDS}}(\mathbf{X}_t, T) = w(t)(\epsilon_\phi(\mathbf{X}_t, t, T) - \epsilon), \quad (1)$$

where $w(t)$ is a weighting function. The gradient direction generated on all the rendered views is used to update θ to produce images that match the conditioned text prompt under diffusion prior. We also follow the SJC [41] to apply the perturb and average scoring into the SDS process. Please refer to [28, 41] for the complete details.

3.3. Editable 3D Scene Layout

The 3D scene layout explicitly combines language structures with 3D layouts in an editable way. Given the input text prompt T , the attribute-object pairs can be easily obtained based on user control. Note that the text prompt indicates the multi-object text prompt by default. As shown in Fig. 3, we can extract multiple noun phrases with their binding attributes and map these local text prompts into corresponding regions. Specifically, we define the scene structure with m local frames, each employs a local NeRF θ_l as representation, the local text prompt $T_l \subseteq T$ and its spatial layout with 3D boxes $\mathbf{b} = \{\mathbf{p}, \mathbf{s}\} \in \mathbb{R}^6$ of each object entity, where $\mathbf{p} = \{p_x, p_y, p_z\}$ refers to the center point and $\mathbf{s} = \{s_x, s_y, s_z\}$ denotes the box scale. *Our editable 3D layout is easy to be collected and edited with its simplicity, allowing for versatile and interactive user control by modifying the box’s or text’s properties to define a new scene.* Moreover, as depicted in Fig. 1, each component in a 3D scene layout can be replaced or re-composed with other trained local NeRFs, which is more friendly for flexible user editions compared with using only text prompts.

3.4. Scene Rendering Pipeline

In CompoNeRF, the scene images are rendered by a ray-casting approach following the design of NeRF. The camera is defined by a pinhole camera model, casting a set of rays $(\mathbf{r}_o, \phi_d, \theta_d) = \mathbf{o} + t\mathbf{d}$ through each pixel on the frame of size $H \times W$, where the $\mathbf{r}_o \in \mathbb{R}^3$ is the origin and the (ϕ_d, θ_d) is the viewing direction. Along this ray, we sample all the points intersected with any layout box of local frames. For each hit sampled point, the color and volumetric density are computed through the local NeRF of the hit local frame. The ray color perdition is calculated by the

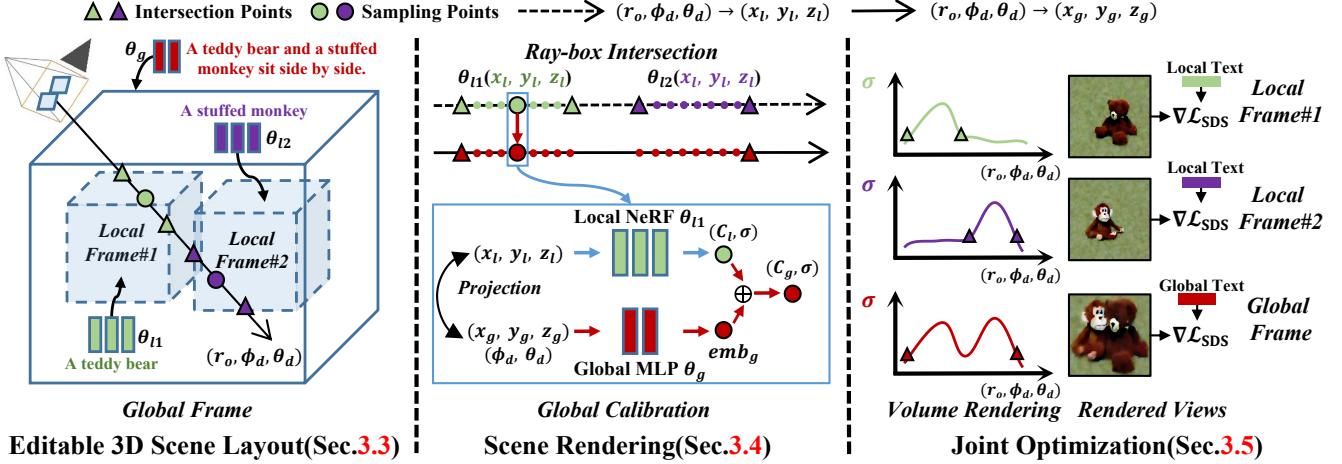


Figure 3: Framework Overview. CompoNeRF consists of three parts: 1). The editable 3D scene layout configures the scene representations with 3D boxes and text prompts; 2). The scene rendering includes the global calibration and the compositional process; 3). The joint optimization applies global and local text guidance on global and local render views.

differentiable integration applied on all the point-predicted colors and volumetric density along the ray.

Ray-box Intersection with Local Frames. Given a ray r_i , each box b_j of the local frame is applied with the AABB ray intersection test algorithm to check the intersections. When the ray r_i is hit with a box b_j of the local frame, we use the entrance and exit points as near t_{in} and far t_{out} bounds to sample N equidistant quadrature points, $t_{i,j,n} = \frac{n-1}{N-1} (t_{out} - t_{in}) + t_{in}, n \in [1, N]$. Note that the coordinates of sampled points are first projected into normalized coordinates using the box scale of local frames to enable each local NeRF to learn the scale-independent representation. The bounding box \mathbf{b} of the local frame in global coordinate can be transformed into a canonical bounding box by $(\mathbf{b} - \mathbf{p})/\mathbf{s}$. Considering the rendering efficiency, we only calculate the valid points, interacted with the boxes, and set all the empty points with a constant background color.

Global Calibration. The 3D boxes are only used for the spatial configuration of local NeRFs, while the implicit representation of local NeRFs is inferred by the canonical samples inside the local frame without considering the global relationship across different objects. To relieve such location-dependent effects, we further calibrate the output color and density from the local NeRF with global coordinates (x_g, y_g, z_g) and ray directions (ϕ_d, θ_d) as the conditional input. Specifically, we adopt a shared MLP θ_g to calibrate all the predicted object colors, that is,

$$C_g = C_l + emb_g = C_l + \theta_g(x_g, y_g, z_g, \phi_d, \theta_d), \quad (2)$$

where C_l is the color predicted by the local NeRF. Therefore, the scene color can preserve the view-consistent behavior from the original architecture and add consistency across poses for the volumetric density. Since the color and density values share the same latent expression in

(x_l, y_l, z_l) , we only calibrate the emitted scene color explicitly with the scene location, as the densities of local NeRFs also are implicitly adjusted during optimization.

Global and Local Volumetric Rendering. After compositing all the interacted points, each ray r_i collects a set sampling points by $\{t_{i,j,n}\}_{j=1,n=1}^{m_j,N}$, where m_j is the number of the hit object. For each sampling point, the inference results with the respective 3D representations are the local color c_l , global color c_g , and density σ . Considering the object occlusions along the ray, we sort the predicted results according to the depth values, *i.e.*, the distances to the camera. Then, the global color \hat{C}_g of ray is calculated by the volumetric rendering equation,

$$\hat{C}_g(r) = \sum_{k=1}^{m_j * N} T_k (1 - \exp(-\sigma_k \delta_k)) C_{g,k}, \quad (3)$$

where $T_k = \exp\left(-\sum_{d=1}^{k-1} \sigma_d \delta_d\right),$

where δ is the distance between adjacent sampled points. For each local NeRF θ_j , we also render the local color \hat{C}_l of a hit ray r_i from sampled points $\{t_{i,j,n}\}_{n=1}^N$ by,

$$\hat{C}_l(r) = \sum_{k=1}^N T_k (1 - \exp(-\sigma_k \delta_k)) C_{l,k}. \quad (4)$$

In fact, each local frame only has a small number of hit rays compared to the scene. Despite the fact that parts of rays are skipped, we observe that it is enough to represent each object accurately while maintaining short rendering times.

3.5. Joint Optimization

For each scene described by the multi-object text prompt T , we optimize m local NeRFs $\{\theta_l\}_{l=1}^m$ according to the layout boxes and a global MLP θ_g for global calibration. To enhance the guidance of local representations, we use

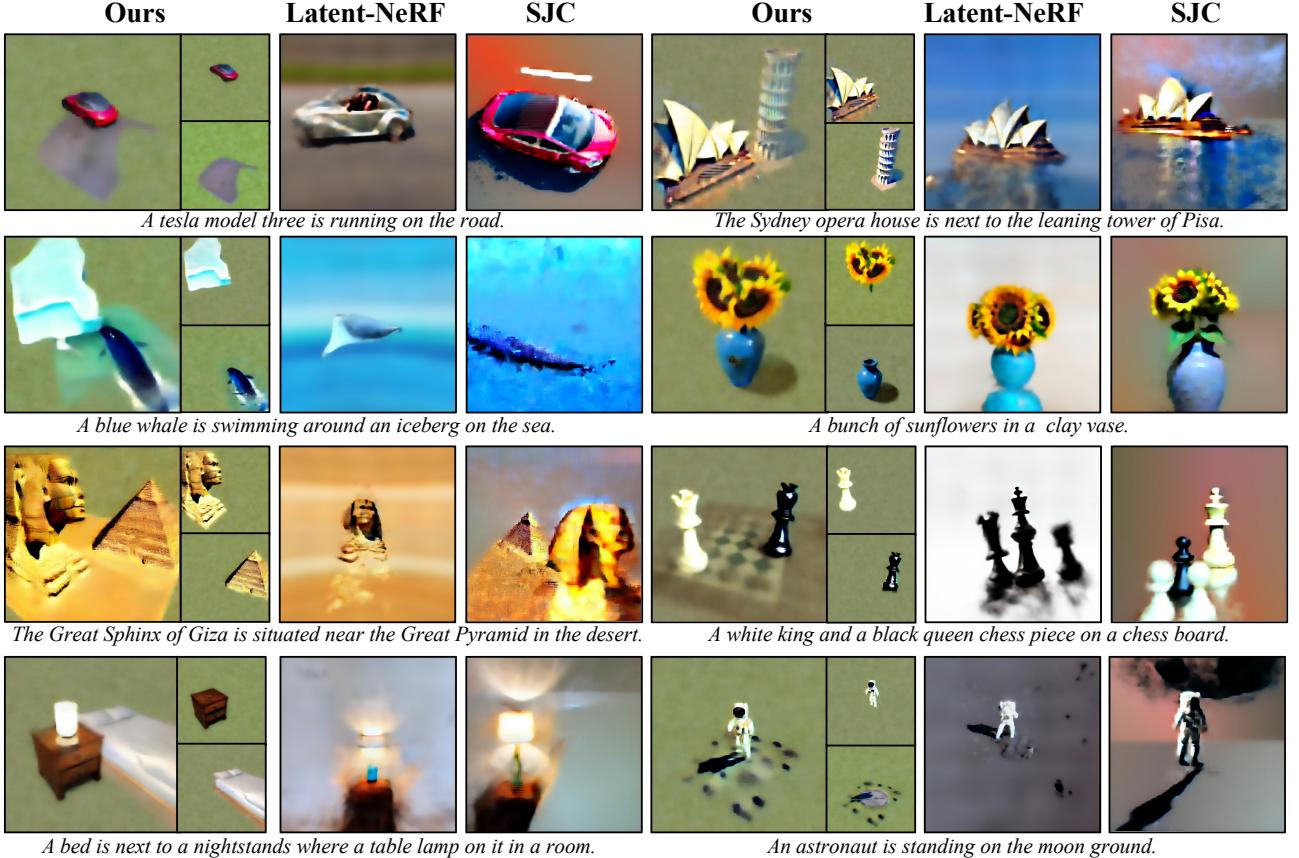


Figure 4: Qualitative comparison with other text-to-3D methods using multi-object text prompts.

the local text prompt $T_l \subseteq T$ of single object to optimize the local NeRFs in local views, as shown in Fig. 3. The scene views $\hat{\mathbf{X}}_g = \{\hat{\mathbf{C}}_{g,i}\}_{i=1}^{H \times W}$ is obtained from the predicted pixel values of $H \times W$ rays by compositing all the ray-box interaction values. Similarly, the rendered view $\hat{\mathbf{X}}_{l,j}$ of the local frame θ_j without compositing other objects can be calculated by $\hat{\mathbf{C}}_{l,j}$, as depicted in Sec. 3.4. We use the local color instead of the globally calibrated color to obtain a local view because the local NeRF should learn the object identity unrelated to its placed position, as the position can be different during user edition. Formally, we employ the following loss as the learning objective,

$$\mathcal{L} = \alpha_g \nabla \mathcal{L}_{\text{SDS}}(\hat{\mathbf{X}}_g, T) + \alpha_l \sum_{j=1}^m \nabla \mathcal{L}_{\text{SDS}}(\hat{\mathbf{X}}_{l,j}, T_{l,j}) + \beta \mathcal{L}_{\text{sparse}},$$

where T denotes the global text prompt, the local text prompt T_l is a subset of T with only the single object. The α_g , α_l , and β are the hyperparameters of the loss weights. $\nabla \mathcal{L}_{\text{SDS}}$ is the score distillation sampling loss, as described in Sec. 3.2. $\mathcal{L}_{\text{sparse}}$ suggested in [19] penalizes the binary entropy of local NeRF density to reduce the floating radiance clouds. For the different levels of text prompts, we both add the directional text prompt (*e.g.*, “front view”, “side view” regarding the global camera pose during training) to the in-

put text prompt similar to [28, 19].

4. Experiments

4.1. Implementation Details.

For score distillation sampling, we use the v1-4 checkpoint of Stable Diffusion based on latent diffusion model [31]. For 3D representation, we use the codebase provided by [19], with the grid encoder from Instant-NGP [24] as our NeRF model and a global MLP consists 6 Linear layers with 64 hidden channels. In the training loss, we set $\alpha_g = 100$, $\alpha_l = 50$, and $\beta = 5e^{-4}$. Our 3D scenes are optimized with a batch size of 1 using the Adam [12] optimizer on a single RTX3090. Please refer to appendix for more details.

4.2. Qualitative Comparison.

In Fig. 4, we show qualitative comparisons of generated 3D assets given the same multi-object text prompt with the Latent-NeRF [19] and SJC [41], which are the SoTAs based on the same Stable Diffusion model. We observe that our method can accurately generate complex 3D models over a diverse set of prompts with more accurate object identity and more sensible structure than compared methods in the

Text: *An apple and a banana.*

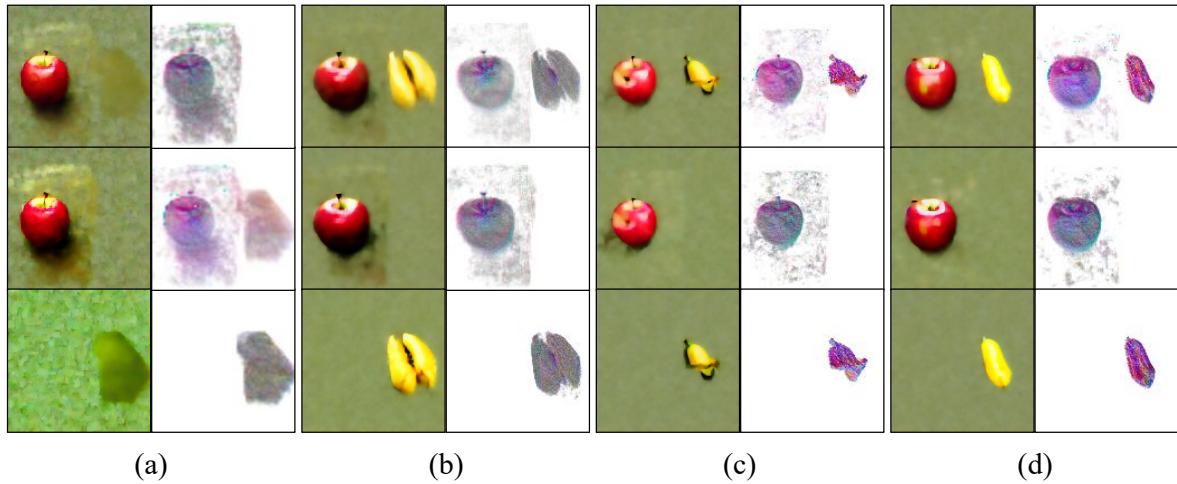


Figure 5: Ablation study on text guidance. (a) without local SDS losses. (b) without global SDS losses. (c) vanilla SDS losses without perturb and average scoring [41]. (d) full model.

Text: *Crystal ball with a wooden base, Dichroic glass ball, Murano glass ball, and Solar-powered glass ball are placed on the glass table.*

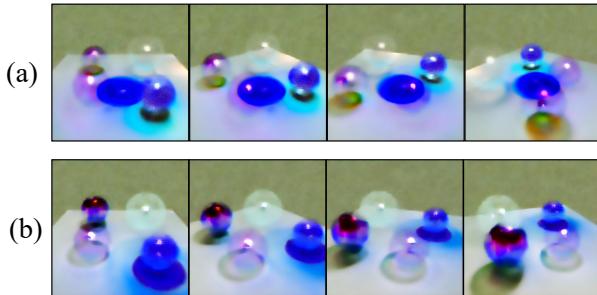


Figure 6: Ablation study on global calibration. (a) without global calibration. (b) full model.

showcases. Note that we cannot validate the predicted results of DreamFusion [28] and Magic3D [15] model as they are built upon on close-sourced diffusion models. Besides, our underlying 3D representation can also be equipped with most object-centric methods [28, 15] once they are released to achieve better single-object modeling, the same as the Latent-NeRF backbone used in our CompoNeRF.

4.3. Ablation Study

Local and Global Text Guidance. We conduct ablations to demonstrate the importance of introducing global and local level guidance discussed in Sec. 3.5. In Fig. 5, we observe that our complete method (Ours) improves the generation accuracy of object identity and better geometry. Note that, without local-level SDS losses, the object frame of 'banana' in Fig. 5(a) even can not be rendered with any details, while without global-level SDS loss, the local frame of 'banana' in Fig. 5(b) can not generate the 'banana' at accurate number. The phenomenon is consistent with our observation

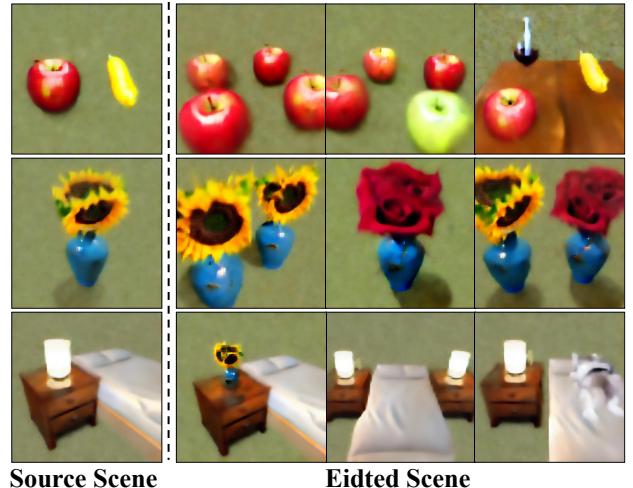


Figure 7: Scene editing results from various types of manipulation on 3D layout, text prompt and scene re-composition

that the generative ability of the pre-trained diffusion model may fail to provide accurate guidance for the multi-object text prompt. In Fig. 5 (c), we also show the perturb and average scoring strategy [41] can generate better geometry results against the vanilla SDS losses [28].

Global Calibration. We further study the global calibration by rendering a scene with glass balls made of different materials in Fig. 5. The rendered results show that with the global calibration procedure, the light reflection and shadow of all balls have more view consistency. While without global calibration, the blue ball renders shadow artifacts that are against scene coherence.

4.4. Editable Scene Rendering and Finetuning.

Due to the compositional capacity brought by the editable 3D scene layout, we can perform scene editing by

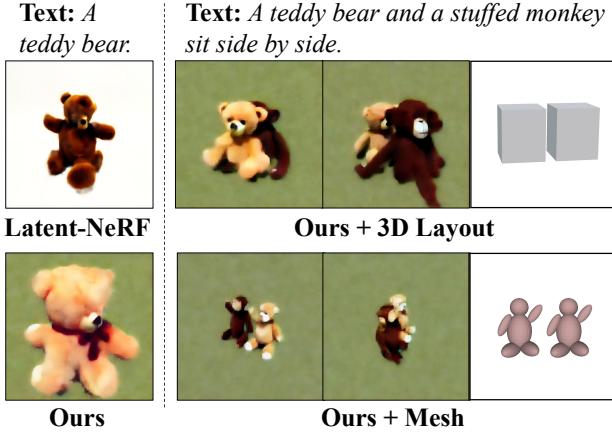


Figure 8: Multi-face problem and more results with different prompt. Note that even for the single object case, all of Latent-NeRF [19], SJC [41], and ours suffer from different extent of multi-face problem.

text editing, moving, scaling, duplicating, removing the single object, and re-compositing a new scene by manipulating the layout of each learned local NeRF. Fig. 7 shows our edited scene rendering results based on a readily well-trained scene. We can see that the manipulated objects are seamlessly integrated into the scene while ensuring the correct spatial relationship following the user control 3D layout. For text editing on a specific object, we simply change a certain part of the text prompt, *e.g.*, ‘a red apple’ to ‘a green apple,’ at both global and local levels and finetune the scene with a few steps. When moving and scaling existing objects, we only need to adjust the box property, such as the center point and box scale. As the duplication, removal, and re-composition, the user can first input the 3D boxes arrangement and then load each box with a local text prompt from a learned local NeRF collection, *e.g.*, copy the single red apple box into four boxes at different locations. Furthermore, all types of manipulation can be combined together to generate a scene with multiple user control inputs. Alternatively, we can further finetune the edited scene with a few finetuning steps to improve the view consistency.

5. Discussion

Our presented CompoNeRF is yet a preliminary step in handling multi-object text for the text-to-3D generation problem. Still, the generation quality is mainly decided by the behaviors of diffusion models. The CompoNeRF also has several limitations related to diffusion guidance.

Multi-face Problem and Stronger Prompt. Similar to most works, Latent-NeRF [19] and SJC [41], the guidance generated from Stable Diffusion may produce a multi-face problem for certain objects as shown in Fig. 8. The diffusion model can not guarantee to generate satisfactory guidance with the desired direction along with the sampling

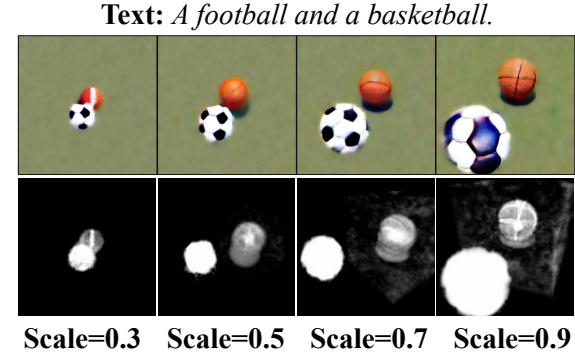


Figure 9: Results of using different training guidance resolutions by scaling the global frames. The first row is the rendering results, and the second row visualizes the rays that hit with local frame for calculating the text guidance.

camera pose. One alternative to relieve the multi-face problem is adding stronger constraints to force the 3D representation to maintain geometric consistency. Our methods also use the mesh constraint, proposed in Latent-NeRF [19], as a more fine-grained 3D layout than the 3D box. Fig. 8 shows that the multi-face problem can be largely relieved with the more accurate mesh constraint. However, the accurate mesh input requires extensive user interaction, which reduces the practical values during application. Nevertheless, we show that our 3D scene layout can be easily extended to more general types of input prompts.

Resolution of Diffusion Model Guidance. Another issue from our framework design is that the guidance resolution of the whole scene and each object identity need to be well balanced. When a single object is placed in a large scene, the rendered view becomes relatively smaller, resulting in a smaller number of pixels that can receive training guidance. Fig. 9 shows the results using the same text prompt while different scales of the global frames vary from 0.3 to 0.9. It indicates that more rays interacting with the local frame can learn a better local NeRF, which is important for large scene rendering as multiple local frames are arranged in the same space while keeping relative size. It is noted that we optimize the scene model in the latent space of the Stable Diffusion model, in which the feature resolution is 64×64 . Therefore, the CompoNeRF needs to trade off the computation efficiency and rendered results quality in the 3D scene layout generation.

6. Conclusion

We proposed a multi-object text-guided compositional 3D scene generation framework, called CompoNeRF, based on an editable 3D scene layout. The 3D scene layout interpreted the multi-object text prompt as a set of local NeRFs binding with a spatial 3D box and object-specific local text prompt. The whole scene view is rendered by

compositing all the local NeRFs defined in the layout. We also designed global calibration and multi-level text guidance mechanisms for improving the quality of the generated 3D scene. Working with the large-scale Stable Diffusion model, we demonstrate that our approach can generate compelling 3D models with multiple objects, comparing favorably to available concurrent work. Finally, we investigate an exciting application of our framework for scene editing and reusing trained models for scene re-composition, identifying an avenue for future work.

References

- [1] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgb-d surface reconstruction. In *CVPR*, pages 6290–6301, 2022. [3](#)
- [2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. [3](#)
- [3] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *ICCV*, pages 5835–5844. IEEE, 2021. [2, 3](#)
- [4] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, pages 5470–5479, 2022. [3](#)
- [5] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022. [2](#)
- [6] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *ICCV*, pages 5712–5721, 2021. [3](#)
- [7] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. 2022. [2](#)
- [8] Michelle Guo, Alireza Fathi, Jiajun Wu, and Thomas Funkhouser. Object-centric neural scene rendering. *arXiv preprint arXiv:2012.08503*, 2020. [3](#)
- [9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. [2](#)
- [10] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535*, 2022. [2](#)
- [11] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *CVPR*, pages 867–876, 2022. [2, 3](#)
- [12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. [6](#)
- [13] Han-Hung Lee and Angel X Chang. Understanding pure clip guidance for voxel grid nerf models. *arXiv preprint arXiv:2209.15172*, 2022. [2](#)
- [14] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900. PMLR, 2022. [2](#)
- [15] 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. *arXiv preprint arXiv:2211.10440*, 2022. [2, 3, 7](#)
- [16] David B Lindell, Julien NP Martel, and Gordon Wetzstein. Autoint: Automatic integration for fast neural volume rendering. In *CVPR*, pages 14556–14565, 2021. [3](#)
- [17] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *ArXiv*, abs/2007.11571, 2020. [3](#)
- [18] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Transactions on Graphics (TOG)*, 40(6):1–16, 2021. [3](#)
- [19] 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. [1, 2, 3, 4, 6, 8](#)
- [20] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. [2](#)
- [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*, volume 12346, pages 405–421. Springer, 2020. [3](#)
- [22] Ashkan Mirzaei, Yash Kant, Jonathan Kelly, and Igor Gilitschenski. Laterf: Label and text driven object radiance fields. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, pages 20–36. Springer, 2022. [2, 3](#)
- [23] 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*, pages 1–8, 2022. [2, 3](#)
- [24] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv preprint arXiv:2201.05989*, 2022. [2, 3, 6](#)
- [25] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, pages 8162–8171. PMLR, 2021. [2](#)
- [26] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *CVPR*, pages 11453–11464, 2021. [3](#)

- [27] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In *CVPR*, pages 2856–2865, 2021. 3
- [28] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2, 3, 4, 6, 7
- [29] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *CVPR*, pages 10318–10327, 2021. 3
- [30] 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 *ICML*, pages 8748–8763. PMLR, 2021. 2, 3
- [31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2, 3, 4, 6
- [32] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 3
- [33] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*. 3
- [34] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *CVPR*, pages 18603–18613, 2022. 2, 3
- [35] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 2
- [36] Yeji Song, Chaerin Kong, Seoyoung Lee, Nojun Kwak, and Joonseok Lee. Towards efficient neural scene graphs by learning consistency fields. *arXiv preprint arXiv:2210.04127*, 2022. 3
- [37] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *CVPR*, pages 7495–7504, 2021. 3
- [38] Edgar Tretschek, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *ICCV*, pages 12959–12970, 2021. 3
- [39] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd E. Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *CVPR*, pages 5481–5490. IEEE, 2022. 3
- [40] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *CVPR*, pages 3835–3844, 2022. 3
- [41] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. *arXiv preprint arXiv:2212.00774*, 2022. 1, 2, 3, 4, 6, 7, 8
- [42] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *Advances in Neural Information Processing Systems*, 34:27171–27183, 2021. 3
- [43] Qianyi Wu, Xian Liu, Yuedong Chen, Kejie Li, Chuanxia Zheng, Jianfei Cai, and Jianmin Zheng. Object-compositional neural implicit surfaces. In *Computer Vision-ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*, pages 197–213. Springer, 2022. 3
- [44] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *CVPR*, pages 9421–9431, 2021. 3
- [45] Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. *arXiv preprint arXiv:2212.14704*, 2022. 2, 3
- [46] Yinghao Xu, Menglei Chai, Zifan Shi, Sida Peng, Ivan Skorokhodov, Aliaksandr Siarohin, Ceyuan Yang, Yujun Shen, Hsin-Ying Lee, Bolei Zhou, et al. Discoscene: Spatially disentangled generative radiance fields for controllable 3d-aware scene synthesis. *arXiv preprint arXiv:2212.11984*, 2022. 3
- [47] Bangbang Yang, Yinda Zhang, Yinghao Xu, Yijin Li, Han Zhou, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Learning object-compositional neural radiance field for editable scene rendering. In *ICCV*, pages 13779–13788, 2021. 3
- [48] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv: CVPR*, 2020. 3
- [49] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (TOG)*, 40(6):1–18, 2021. 3
- [50] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *ICCV*, pages 15838–15847, 2021. 3