

GaussianEditor: Editing 3D Gaussians Delicately with Text Instructions

Jiemin Fang*, Junjie Wang*, Xiaopeng Zhang, Lingxi Xie, Qi Tian
Huawei Inc.

{jaminfong, is.wangjunjie, 198808xc, zxphistory}@gmail.com tian.qi1@huawei.com
GaussianEditor.github.io

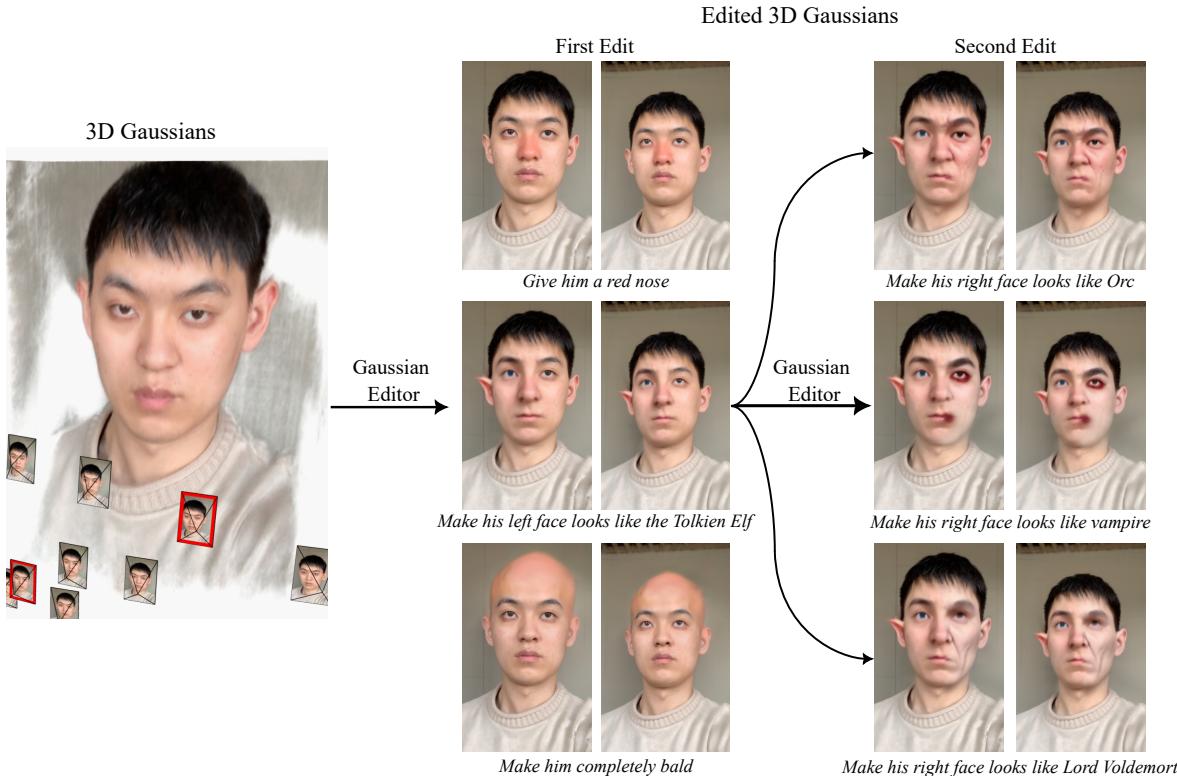


Figure 1. We propose GaussianEditor, an interactive framework to achieve delicate 3D scene editing following text instructions. As shown in this figure, our method can precisely control the editing region and achieve multi-round editing.

Abstract

Recently, impressive results have been achieved in 3D scene editing with text instructions based on a 2D diffusion model. However, current diffusion models primarily generate images by predicting noise in the latent space, and the editing is usually applied to the whole image, which makes it challenging to perform delicate, especially localized, editing for 3D scenes. Inspired by recent 3D Gaussian splatting, we propose a systematic framework, named GaussianEditor, to edit 3D scenes delicately via 3D Gaussians with text instructions. Benefiting from the explicit property of 3D Gaussians, we design a series of techniques to

achieve delicate editing. Specifically, we first extract the region of interest (RoI) corresponding to the text instruction, aligning it to 3D Gaussians. The Gaussian RoI is further used to control the editing process. Our framework can achieve more delicate and precise editing of 3D scenes than previous methods while enjoying much faster training speed, i.e. within 20 minutes on a single V100 GPU, more than twice as fast as Instruct-NeRF2NeRF (45 minutes – 2 hours)¹.

*Equal contributions.

¹The editing time varies in different scenes according to the scene structure complexity.

1. Introduction

Creating 3D assets has played a critical role in many applications and industries, *e.g.* movie/game production, artistic creation, AR, VR *etc.* However, this process is usually expensive and cumbersome, especially for traditional pipelines. Designers need to take a lot of labor and time to finish each step, *e.g.* sketching, building structures, creating textures *etc.* One cheap and effective way of creating high-quality 3D assets is to start from an existing scene, capturing, modeling, and editing the scene and obtaining the wanted one. This approach can be also used for user-interactive entertainment applications.

Neural radiance field methods [2, 3, 6, 29, 31, 43, 46] have shown great power in representing 3D scenes and synthesizing novel-view images. Past years have witnessed the rapid development of NeRF and its variants, from both quality and efficiency perspectives. Editing a pre-trained NeRF model becomes a promising way to effectively edit 3D scenes. Represented by Instruct-NeRF2NeRF [11], researchers propose to use the image-conditioned 2D diffusion model, *e.g.* InstructPix2Pix [4], to edit 3D scenes simply with text instructions. Notable results have been achieved as real scenes can be changed following the given text instruction. However, as current 2D diffusion models have the limited ability to localize editing regions, it is hard to generate delicate edited scenes as the unwanted regions are usually changed with diffusion models. Even though some works [30] propose to constrain the editing region on the edited 2D images, the editing region is still hard to accurately localize and apply to the 3D representation. Besides, NeRF-based methods [9, 49] bear coupling effects between different spatial positions, *e.g.* different points are queried from the same MLP field (for implicit representations) or voxel vertices (for explicit representations).

Recent 3D Gaussian Splatting [18] (3D-GS) has been a groundbreaking work in the field of radiance field, which is the first to achieve a real sense of real-time rendering while enjoying high rendering quality and training speed. Besides its efficiency, we further notice its natural explicit property. 3D-GS has a great advantage for editing tasks as each 3D Gaussian exists individually. It is easy to edit 3D scenes by directly manipulating 3D Gaussians with desired constraints applied.

Aiming at editing 3D scenes delicately, we propose to represent the scene with 3D Gaussians, which can be edited with text instructions, and name our method as GaussianEditor. To achieve precise control for editing regions, we divide our method into three main parts. The first one is the region of interest (RoI) extraction from the given text instruction. The instruction may be complex or indirect while this module helps extract the key words matching the RoI for editing. The second part aligns the extracted text RoI to the 3D Gaussian space through the image space, where

a grounding segmentation model can be applied. The last part is to edit the original 3D Gaussians delicately with constraints in the obtained 3D Gaussian RoI. With the above processes, the region for editing can be precisely localized simply from text instructions, which are used for constraining the 3D Gaussian updating to obtain a delicately edited new 3D scene. Besides, we enable interfaces for users to introduce more exact instructions for more delicate editing, *e.g.* Gaussian point selecting and 3D boxes for modifying the editing regions².

Our contributions can be summarized as follows.

- As far as we know, our GaussianEditor is one of the first systematic methods to achieve delicate 3D scene editing based on 3D Gaussian splatting.
- A series of techniques are designed and proposed to precisely localize the editing region of interest, which are aligned and applied to 3D Gaussians. Though some sub-modules are from existing works, we believe integrating these awesome techniques to work effectively is a valuable topic, which is what we focus on in this paper.
- Our method achieves a series of more delicate editing results compared with the previous representative work Instruct-NeRF2NeRF [11] with much shorter training time (within 20 minutes *v.s.* 45 minutes – 2 hours).

2. Related Work

2D Image Editing with Diffusion Models By virtue of advancements in diffusion model technology [13, 44], numerous generative models [41] have attained remarkably impressive outcomes in the realm of image synthesis. Recent progressions in diffusion models have showcased their adeptness in fabricating lifelike images from arbitrary textual inputs [8, 14, 40, 42, 45]. Harnessing the robust semantic comprehension and image generation capabilities of foundational diffusion models, an escalating number of research explorations are currently employing diffusion models as a fundamental framework for implementing text-based image editing functionalities [33, 37, 38, 41]. Some of these methodologies necessitate the manual provision of captions for both the original and edited images [12], while others mandate specific scenario-based training for optimization [39]. These requisites have rendered it arduous for ordinary users to avail themselves of such techniques. Expanding upon this foundation, iP2P [4] introduces instruction-based capabilities to image editing, enabling users to simply input an image and apprise the model of the desired alterations. This user-friendly approach facilitates the democratization of image editing in a more accessible manner.

²These additional instructions are applied to generate the man with two different edited half faces in Fig. 1.

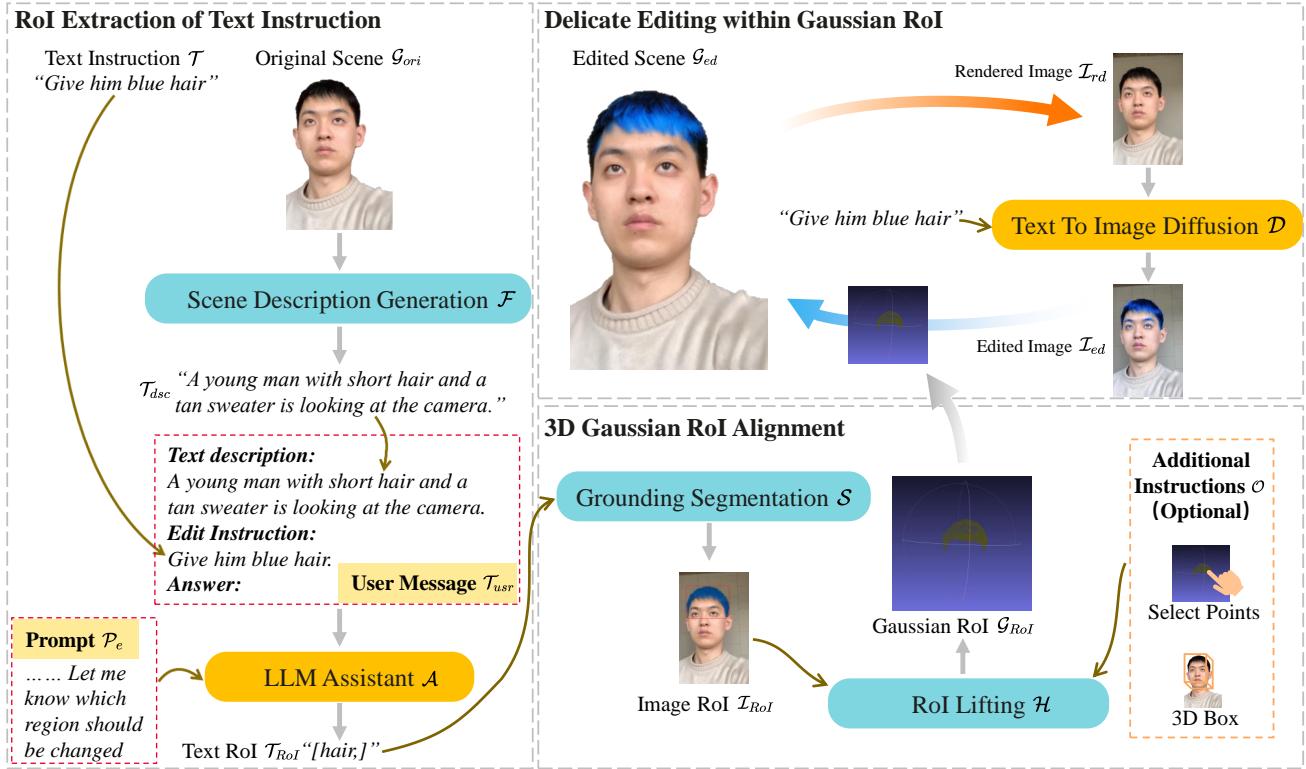


Figure 2. Our framework, named GaussianEditor, consists of three key steps. First, the scene generation \mathcal{F} was performed to get the scene description \mathcal{T}_{desc} from original 3D Gaussians \mathcal{G}_{ori} . The description \mathcal{T}_{desc} and the text instruction \mathcal{T} provided by the user are then combined using a template \mathcal{T}_{tmp} to get user message \mathcal{T}_{usr} . The \mathcal{T}_{usr} and a pre-defined prompt \mathcal{P}_e are fed into an LLM assistant \mathcal{A} to obtain the instruction ROI \mathcal{T}_{roi} . Second, a grounding segmentation model \mathcal{S} is used to convert \mathcal{T}_{roi} to image ROI \mathcal{I}_{roi} , which is then lifted to 3D Gaussians ROI \mathcal{G}_{roi} by ROI lifting \mathcal{H} , where additional user instructions \mathcal{O} can be incorporated. Third, following the user instruction \mathcal{T} , rendered image \mathcal{I}_{rd} from randomly chosen views is edited by a 2D diffusion model \mathcal{D} . The loss between \mathcal{I}_{rd} and edited one \mathcal{I}_{ed} is calculated. Finally, gradient backpropagation and optimization are performed within the Gaussian ROI \mathcal{G}_{roi} to get the edited scene \mathcal{G}_{ed} .

3D Scene Editing of Radiance Fields 3D Scene Editing of Radiance Fields has gained significant popularity as a recent research direction [1, 10, 15, 20, 22–25, 28, 34, 47, 48, 52–54]. The main objective of such methods is to enable the manipulation of both the geometry and appearance of 3D scene representations. However, editing such scenes poses inherent challenges due to the implicit nature of traditional NeRF representations, which lack precise localization capabilities. As a result, previous works have primarily focused on achieving global style transformations of 3D scenes [7, 16, 17, 32, 49, 51, 56]. While some efforts have been made towards object-centric scene editing [57], keeping the background unchanged has been a persistent challenge. For example, a recently proposed work, Instruct-NeRF2NeRF [11], implements text instruction-controlled 3D scene editing capabilities, achieving excellent editing effects while maintaining user-friendliness. However, it relies on the editing effect of 2D images, which may cause global changes to the 3D scene. A subsequent work [30] attempts to compute the relevance map between edited and unedited images to localize the editing area. The relevance

map may be unreliable when the 2D IP2P [4] model fails. Other efforts [23] attempt to rely on the 3D coordinates entered by the user to decide the editing area. The introduction of 3D Gaussians [18] has provided an opportunity to address this limitation. Its explicit 3D representation methodology enables accurate selection and manipulation of editing areas. By incorporating LLMs, the whole process can be more automated.

3. Method

In this section, we first review 3D representation methods in Sec. 3.1. Subsequently, in Sec. 3.2, we overview our proposed approach, which mainly includes three modules. Sec. 3.3 delves into the precise Region of Interest (RoI) extraction of text instructions, while Sec. 3.4 introduces how to align the instruction RoI with 3D Gaussians. Finally, Sec. 3.5 describes the delicate editing process within the obtained Gaussian ROI.

3.1. Preliminaries

3D Gaussian Splatting. 3D Gaussian splatting [18] is a recent powerful 3D representation method. It represents the 3D scene with point-like 3D Gaussians $\mathcal{G} = \{g_1, g_2, \dots, g_N\}$, where $g_i = \{\mu, \Sigma, c, \alpha\}$ and $i \in \{1, \dots, N\}$. Among them, $\mu \in \mathbb{R}^3$ is the position where the Gaussian centers, $\Sigma \in \mathbb{R}^7$ denotes the 3D covariance matrix, $c \in \mathbb{R}^3$ is the RGB color and $\alpha \in \mathbb{R}^1$ is the opacity. Benefiting from the compact representation of Gaussians and efficient differentiable rendering approach, 3D Gaussian splatting achieves real-time rendering with high quality. The splatting rendering process can be formulated as

$$C = \sum_{i \in \mathcal{N}} c_i \sigma_i \prod_{j=1}^{i-1} (1 - \alpha_j), \quad (1)$$

where $\sigma_i = \alpha_i e^{-\frac{1}{2}(x_i)^T \Sigma^{-1}(x_i)}$ represents the influence of the Gaussian to the image pixel and x_i is the distance between the pixel and the center of the i -th Gaussian.

3.2. Overall Framework

Given a group of 3D Gaussians \mathcal{G}_{ori} for a scene and a text instruction \mathcal{T} for editing, our Gaussian editor \mathcal{E} can edit the 3D Gaussians delicately into a new one, denoted as \mathcal{G}_{ed} , with the guidance of the instruction. The whole process can be formulated as

$$\mathcal{G}_{ed} = \mathcal{E}(\mathcal{G}_{ori}, \mathcal{T}). \quad (2)$$

Fig. 2 illustrates the overall framework of our approach, which consists of three main steps. First, the region of interest (RoI) is extracted from the text instruction. In this step, we employ a multimodal model \mathcal{C} to generate the scene description \mathcal{T}_{dsc} from the rendered image \mathcal{I}_{rd} of original 3D Gaussians \mathcal{G}_{ori} . We then input the scene description \mathcal{T}_{dsc} and text instruction \mathcal{T} into a large language model assistant \mathcal{A} to determine where we should make edits in the scene. The output of this step is referred to as the instruction Region of Interest (RoI) \mathcal{T}_{RoI} .

The next step is the 3D Gaussian RoI alignment. We use a grounding segmentation model \mathcal{S} to convert the RoI from text space, *i.e.* \mathcal{T}_{RoI} , to the image space, *i.e.* \mathcal{I}_{RoI} . Then the image RoI \mathcal{I}_{RoI} is lifted to the RoI of 3D Gaussians \mathcal{G}_{RoI} through training. The Gaussian RoI allows us to precisely control the regions where edits will be applied.

The last step is delicate editing within the Gaussian RoI. In this step, we randomly sample the view to obtain the rendered image \mathcal{I}_{rd} . A 2D diffusion model \mathcal{D} is used to perform the editing process on the rendered image \mathcal{I}_{rd} , with the user instruction \mathcal{T} and the original image \mathcal{I}_{ori} as conditions. The resulting edited image is denoted as \mathcal{I}_{ed} . Subsequently, we calculate the loss between \mathcal{I}_{ed} and \mathcal{I}_{rd} and make gradient back-propagation within \mathcal{G}_{RoI} . This implies that only

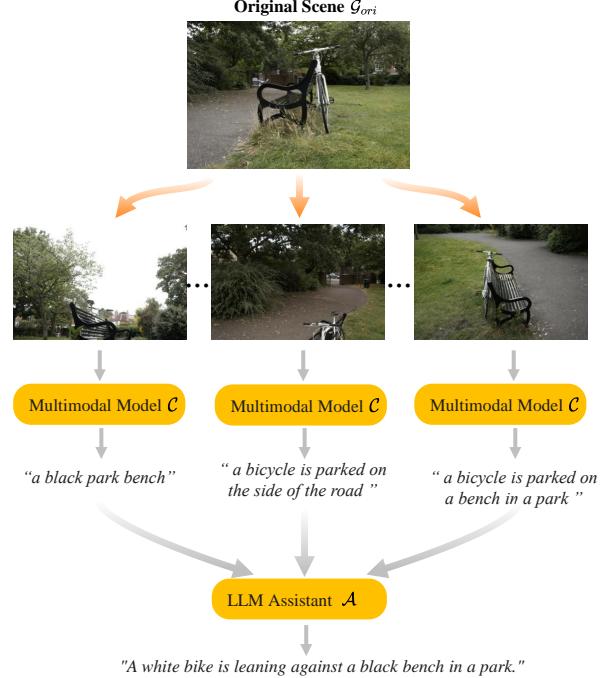


Figure 3. The process of obtaining scene description.

the regions specified by the RoI can receive corresponding gradients during the back-propagation process. Finally, optimization is executed based on these gradients. Through several rounds of iterative optimization, the final optimized scene representation \mathcal{G}_{ed} is obtained.

3.3. RoI Extraction of Text Instruction

The instruction RoI is extracted for the editing regions from both the 3D scene representation \mathcal{G}_{ori} and the text instruction \mathcal{T} provided by the user. To achieve this, we employ a multimodal model \mathcal{C} in conjunction with the large language model assistant \mathcal{A} . The first step is the scene description generation \mathcal{F} , which aims to get the text description \mathcal{T}_{dsc} from 3D Gaussians \mathcal{G}_{ori} :

$$\mathcal{T}_{dsc} = \mathcal{F}(\mathcal{G}_{ori}). \quad (3)$$

By leveraging the technique of differentiable splatting as shown in Eq. 1, a set of 2D image samples \mathcal{I}_{rd} are generated, the quantity of which is recorded as \mathcal{M} . These sampled images $\mathcal{I}_{sp} = \{\mathcal{I}_1, \mathcal{I}_2, \dots, \mathcal{I}_{\mathcal{M}}\}$, are then inputted into a multimodal model \mathcal{C} to generate corresponding text descriptions $\mathcal{T}_{img} = \{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_{\mathcal{M}}\}$:

$$\mathcal{T}_k = \mathcal{C}(\mathcal{P}_c, \mathcal{I}_k), k \in \mathcal{M}, \quad (4)$$

where \mathcal{P}_c is a prompt given by the user, such as “What is the content of the image”, to multimodal model \mathcal{C} for getting precise description. Subsequently, these descriptions are fed into a large language model, which works as a scene understanding assistant \mathcal{A} . The assistant \mathcal{A} is

specifically instructed by a prompt \mathcal{P}_m , to merge descriptions of diverse views into one detailed scene description \mathcal{T}_{dsc} :

$$\mathcal{T}_{dsc} = \mathcal{A}(\mathcal{P}_m, \mathcal{T}_{img}). \quad (5)$$

The process of the scene description generation \mathcal{F} is shown in Fig. 3.

Next, large language model assistant \mathcal{A} is used to extract the instruction RoI. Initially, we set a prompt \mathcal{P}_e to instruct \mathcal{A} about the task for assisting in instruction RoI extraction. Then the scene description \mathcal{T}_{dsc} and the user instruction \mathcal{T} are combined with a predefined template \mathcal{T}_{tmp} : “Text description: \mathcal{T}_{dsc} Edit Instruction: \mathcal{T} Answer:” to form the user message $\mathcal{T}_{usr} = \mathcal{T}_{tmp}(\mathcal{T}_{dsc}, \mathcal{T})$. During the actual execution, we input both the prompt \mathcal{P}_e and the user message \mathcal{T}_{usr} into \mathcal{A} to obtain the instruction RoI \mathcal{T}_{RoI} :

$$\mathcal{T}_{RoI} = \mathcal{A}(\mathcal{P}_e, \mathcal{T}_{usr}). \quad (6)$$

3.4. 3D Gaussian RoI Alignment

To confine the 3D editing region within the instruction RoI, 3D Gaussian RoI \mathcal{G}_{RoI} is aligned with the text RoI \mathcal{T}_{RoI} in this step. First, The RoI in the text space is transformed into the image space via a grounding segmentation model \mathcal{S} :

$$\mathcal{I}_{RoI} = \mathcal{S}(\mathcal{I}_{rd}, \mathcal{T}_{RoI}), \quad (7)$$

where \mathcal{I}_{rd} is rendered image of the original scene \mathcal{G}_{ori} .

Then we lift the RoI \mathcal{T}_{RoI} in the image space to 3D Gaussian \mathcal{G}_{RoI} through training. To achieve this, an additional RoI attribute $r \in \mathbb{R}^1$ was added to 3D Gaussian $g_i = \{\mu_i, \Sigma_i, c_i, \alpha_i, r_i\}$. r is initialized to 0, which means it is not in the Gaussians RoI, and 1 means it is inside the RoI. The set of r is denoted as $\mathcal{R} \in \mathbb{R}^{N, 1}$, where the N is the number of 3D Gaussians \mathcal{G}_{ori} .

Then the color c_i in Eq. 1 was rewritten with r_i to get the rendered RoI \mathcal{I}_{RoI}^{rd} :

$$\mathcal{I}_{RoI}^{rd} = \sum_{i \in N} r_i \sigma_i \prod_{j=1}^{i-1} (1 - \alpha_j). \quad (8)$$

Taking inspiration from SA3D [5], to get the trained Gaussians RoI \mathcal{G}_{RoI}^{tr} , we adopt a similar loss function to supervise the training process:

$$\mathcal{L}_{proj} = \lambda_1 \sum (\mathcal{I}_{RoI}^{rd} \cdot \mathcal{I}_{RoI}) + \lambda_2 \sum ((1 - \mathcal{I}_{RoI}^{rd}) \cdot \mathcal{I}_{RoI}), \quad (9)$$

where λ_1 and λ_2 are hyperparameters. Additionally, the user can modify the trained Gaussian RoI \mathcal{G}_{RoI}^{tr} by giving added Gaussian RoI \mathcal{G}_{RoI}^{add} , deleted Gaussian RoI \mathcal{G}_{RoI}^{del} and 3D box \mathcal{B}_{3D} :

$$\mathcal{G}_{RoI} = (\mathcal{G}_{RoI}^{tr} \cup \mathcal{G}_{RoI}^{add} - \mathcal{G}_{RoI}^{del}) \cap \mathcal{B}_{3D}, \quad (10)$$

\mathcal{G}_{RoI}^{add} represents the 3D Gaussians user wants to edit, \mathcal{G}_{RoI}^{del} means 3D Gaussians user wants to keep from editing, \mathcal{B}_{3D} is the coordinates of 3D cuboid it limits RoI to inside the box. \mathcal{G}_{RoI} is the aligned RoI with the text RoI. For example, when editing the left face of the man in Fig. 1, grounding segmentation failed to ground “left face”, instead, it grounded the whole face. In this scenario, the user can use the interactive interface to set the right face as \mathcal{G}_{RoI}^{del} or enter the rectangular box where the left face is located as \mathcal{B}_{3D} . The lifting process \mathcal{H} can be represented as:

$$\mathcal{G}_{RoI} = \mathcal{H}(\mathcal{I}_{RoI}, \mathcal{O}), \quad (11)$$

where $\mathcal{O} = \{\mathcal{G}_{RoI}^{add}, \mathcal{G}_{RoI}^{del}, \mathcal{B}_{3D}\}$ is additional user instructions.

3.5. Delicate Editing within Gaussian RoI

To achieve delicate editing in 3D scenes, we use the Gaussian RoI to constrain the editing area. In particular, we randomly sample viewpoints from the 3D scene representation and render 2D image representations \mathcal{I}_{rd} . Then 2D diffusion model \mathcal{D} is used to edit the rendered image \mathcal{I}_{rd} , with the user instruction \mathcal{T} and original image \mathcal{I}_{ori} as conditions, to get the edited image \mathcal{I}_{ed} as

$$\mathcal{I}_{ed} = \mathcal{D}(\mathcal{I}_{rd}, t; \mathcal{T}, \mathcal{I}_{ori}), \quad (12)$$

where t is a randomly chosen noise level from $[t_{min}, t_{max}]$.

Similar to 3D-GS [18], we apply the \mathcal{L}_1 and D-SSIM loss functions during editing.

$$\mathcal{L} = (1 - \beta)\mathcal{L}_1 + \beta\mathcal{L}_{D-SSIM}. \quad (13)$$

the two losses are calculated between the 2D edited image \mathcal{I}_{ed} and the rendered image \mathcal{I}_{rd} . Then, gradient backpropagation is performed within Gaussian RoI \mathcal{G}_{RoI} :

$$\nabla \mathcal{G} = \frac{\partial \mathcal{L}}{\partial \mathcal{G}} \cdot \mathcal{R}, \quad (14)$$

where \mathcal{R} is the set of RoI attributes. That means only Gaussians in RoI can receive gradients. Finally, we utilize the Adam algorithm to optimize the 3D Gaussians. After it rounds of training, the edited 3D scene \mathcal{G}_{ed} is obtained.

4. Experiments

4.1. Implementation Details

Our method is implemented in PyTorch [35] and CUDA, based on 3D Gaussian splatting. The multimodal model we used in our method is BLIP2 [21], and we use GPT-3.5 Turbo to ground the text ROI. For grounding segmentation, We use the cascade strategy, *i.e.* first using Grounding DINO [26] to get the box on the image corresponding to the text, and then using SAM [19] to get the corresponding image RoI. The 2D diffusion model used in our method is Instruct Pix2Pix [4]. We leave more details in the Appendix.



Figure 4. Qualitative results on outdoor scenes. Our method supports separate foreground and background editing in real-world scenes.

4.2. Qualitative Evaluation

Visualization Results. In Fig. 1 and Fig. 4, we present the visual results of our method, demonstrating the precise editing effects while ensuring 3D consistency. Fig. 1 demonstrates the editing capabilities of our method on human subjects. The first column displays the original scenes. In the second column, the first row “Give him a red nose” illustrates our ability to change colors. The third row in the second column, “Make him completely bald”, showcases our geometric editing capability. The second row in the second column demonstrates our precise editing ability by exclusively editing the left side of the person’s face. Building upon this foundation, we achieve editing in the third column, specifically focusing on the right side of the face, showcasing our ability to perform multi-round edits and accurately fulfill user instructions. Fig. 4 showcases the precise editing capabilities of our method in open 3D scenes. In the upper portion, the bicycle scene allows us to accurately locate the position of the road and edit its texture, transforming it into the grass, a river. It is worth noting that in the experiment where we change the texture to a river, our method accurately constructs the reflection, making it appear realistic. On the basis of editing the road

into a river, we further edited the bench, proving that our method can achieve multiple rounds of editing. The lower portion demonstrates the results of editing the bear using our method, where we fully preserve the original appearance of the background area and focus the edits on the bear.

Comparisons with Instruct-NeRF2NeRF. Fig. 5 compares the results of our method with those of IN2N [11], on the scenes presented in IN2N. From the figure, it is evident that our method successfully changes the texture of pants without affecting the clothes, and vice versa, demonstrating the effectiveness of our method in distinguishing different objects within the foreground. Additionally, when editing the clothes and pants, the background remains unaffected, indicating our method’s effective separation of foreground and background. Furthermore, the last column reveals that IN2N, limited by 2D diffusion, distorts the face, while our method maintains a superior rendering quality of faces.

Complex Multi-Object Scenes. Furthermore, we present the results of our editing in a complex scene featuring multiple objects, as depicted in Fig 6. We specifically selected three distinct object types for editing purposes. The first object type is the background. In a particular sub-scene, the desktop shares similarities with the background. Employing

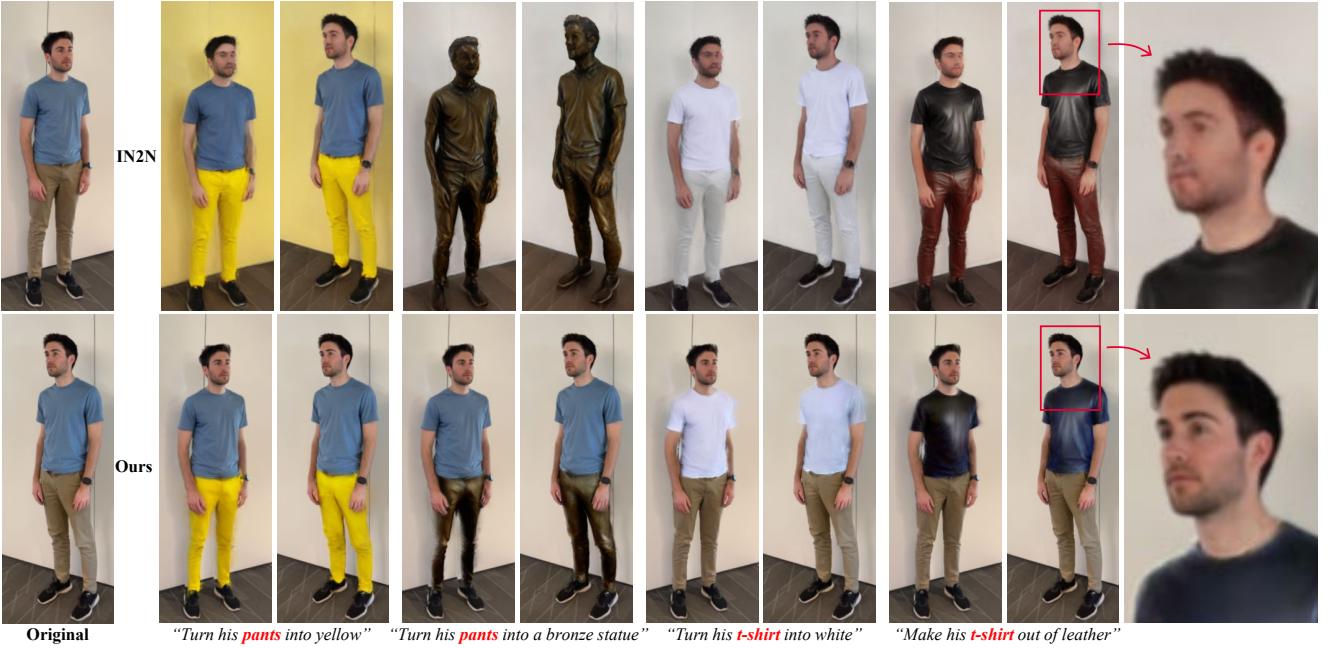


Figure 5. Comparisons with Instruct-NeRF2NeRF (IN2N) [11]. We use IN2N as a baseline and compare our method with the scenes presented in their paper.



Figure 6. Qualitative results on complex multi-object scenes. The background “desk”, the foreground “flower pot”, and the multi-view blocked foreground “rolling pin” are edited separately.

a caption-based approach, we successfully transformed the desktop into a wooden material. As evident in the image, the edited result exhibits a distinct and pronounced wood texture. The second object type is a foreground object. In the sub-scene, the flowerpot serves as one of the foreground objects. We opted to change the color of the flowerpot to red, and the outcome was highly successful. Lastly, the most intricate editing task involved the rolling pin, which was occluded by multiple objects from various perspectives. As illustrated in the lower right corner of the picture, the rolling pin was hindered by different objects, yet we managed to edit it into a cucumber without any adverse impact on the other objects.

4.3. Ablation Study and Analysis

Ablation of Gaussian ROI, Text ROI, ROI Lifting. To validate the effectiveness of each module in our framework, we design three variant approaches: (1) w/o Gaussian ROI: We discontinued the use of Gaussian ROI \mathcal{G}_{ROI} to control the gradients of Gaussian points, as mentioned in Eq. 14.

(2) w/o Text ROI: In this scenario, we ceased the selection of text ROI \mathcal{T}_{ROI} using LLM assistant \mathcal{A} . Instead, all the words in the user’s instruction are treated as text ROI \mathcal{T}_{ROI} . (3) w/o ROI lifting: Instead of lifting the image ROI \mathcal{I}_{ROI} to 3D Gaussians, the image ROI \mathcal{I}_{ROI} is used to govern the calculation of the loss. That is, only the pixels within the image ROI \mathcal{I}_{ROI} are taken into account for the loss computation. Fig. 7 showcases the outcomes of our ablation experiment, which aimed to edit the doll based on the instruction “Turn its mouth red.” The results reveal the following findings. When the Gaussian ROI is not used, the 3D scene is all turned red because the 2D diffusion fails to successfully control the editing area. In cases where text ROI \mathcal{T}_{ROI} is not utilized, the grounding segmentation model tends to segment the entire foreground object, leading to the doll being entirely edited to red. (3) When ROI lifting \mathcal{H} is not employed, the doll’s mouth is successfully turned red, but other facial areas are also affected. We also find that the current 2D grounding segmentation model may fail to

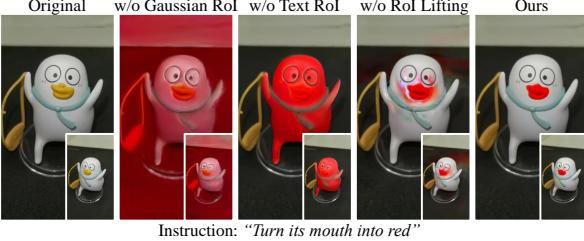


Figure 7. Ablation results of settings without Gaussian RoI, without Text RoI, and without ROI Lifting.

parse specific views, which means there exist noises on the image RoI \mathcal{I}_{RoI} . Consequently, when optimizing the scene based on the \mathcal{I}_{RoI} , leakage occurs during the editing process. Our proposed ROI lifting module effectively addresses this issue during training. In conclusion, our ablation experiment effectively demonstrates the effectiveness of several ROI-related modules in our method.

Ablation of Scene Description Generation. We further conduct experiments to evaluate the role of scene description generation, employing three distinct experimental setups. The first one solely relies on the utilization of the template \mathcal{T}_{tmp} and user instruction \mathcal{T} to compose the user message \mathcal{T}_{usr} , without employing scene description generation. The LLM was employed to acquire the text ROI \mathcal{T}_{RoI} . The second one involves randomly sampling a view and extracting the corresponding image’s text description \mathcal{T}_{img} , which was then incorporated into the template \mathcal{T}_{tmp} as the text description \mathcal{T}_{dsc} and user instructions \mathcal{T} to form the user message \mathcal{T}_{usr} , which can be denoted as $\mathcal{T}_{usr} = \mathcal{T}_{tmp}(\mathcal{T}_{img}, \mathcal{T})$. The third one represents the complete version of our approach. The test scene involves a park where a bike and a bench are positioned closely together. The editing instruction is “Turn the thing next to the bike orange”. The obtained results are presented in Fig. 8. As is evident from the image, when description generation is not employed, the LLM fails to acquire the Text ROI according to the user instructions, resulting in editing failure. The second one randomly samples images to obtain text descriptions, resulting in incomplete descriptions and leading to an incorrect text ROI prediction by the LLM. Consequently, the final editing result turns the road into an orange color. In contrast, our method flawlessly executes the editing task. This success can be attributed to scene description generation, which obtains an accurate text description encompassing the relative positional relationship between the bicycle and the bench. This enables the LLM to analyze and determine the user’s intention to edit the bench. Consequently, the desired color change of the bench is successfully implemented.

4.4. Limitations

Although our framework has solved some problems inherited from the integrated sub-modules, *e.g.* noise in the re-

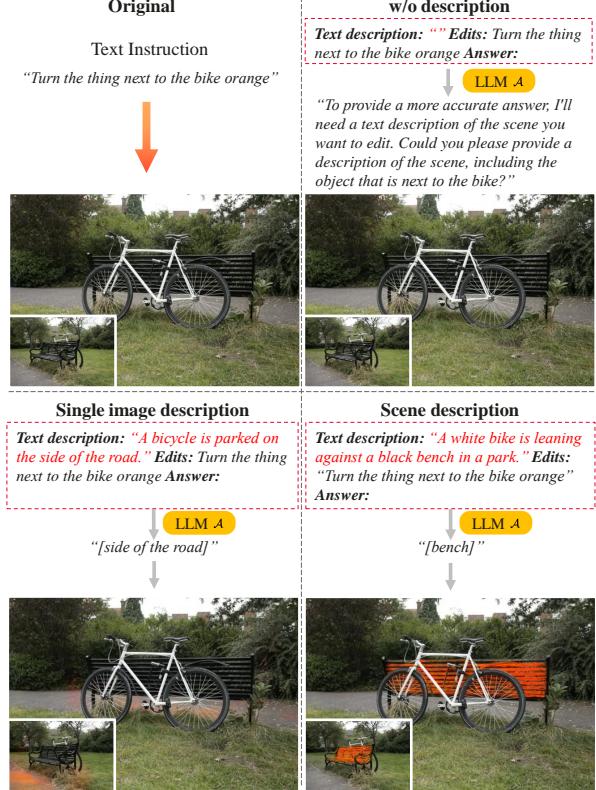


Figure 8. Ablation results about the scene description generation.

sults of grounding segmentation, there are still some problems that the current system cannot completely avoid. In scene description generation, the descriptions from different views of the same object may differ from each other. When the differences are large enough, the LLM may misunderstand these descriptions as those from multiple objects. This issue does not affect the results in the current experiment, but we would like to optimize this in the future. In addition, our system cannot achieve good editing results in scenes where the grounding segmentation or diffusion model completely fails.

5. Conclusion

This paper proposes a systematic framework, named GaussianEditor, for text-guided delicate 3D scene editing. As we know, GaussianEditor is the first work to edit 3D Gaussians, taking advantage of the explicit property of 3D Gaussians and making it easy to control the editing area precisely. Several techniques are proposed to achieve delicate editing, including extracting instruction ROI from texts, aligning the ROI to 3D Gaussians, and editing the scene with the Gaussian ROI. GaussianEditor achieves notably more delicate editing results than IN2N [11] with much shorter training time (within 20 minutes *v.s.* 45 minutes – 2 hours). Noticing recent works [27, 50] have extended Gaussian splatting to dynamic scenes, we leave the delicate editing in dynamic scenes as future work.

References

- [1] Chong Bao, Yinda Zhang, Bangbang Yang, Tianxing Fan, Zesong Yang, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Sine: Semantic-driven image-based nerf editing with prior-guided editing field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3
- [2] 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*, 2021. 2
- [3] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, 2022. 2
- [4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022. 2, 3, 5, 12
- [5] Jiazhong Cen, Zanwei Zhou, Jiemin Fang, Chen Yang, Wei Shen, Lingxi Xie, Dongsheng Jiang, Xiaopeng Zhang, and Qi Tian. Segment anything in 3d with nerfs. In *NeurIPS*, 2023. 5
- [6] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *ECCV*, 2022. 2
- [7] Pei-Ze Chiang, Meng-Shiun Tsai, Hung-Yu Tseng, Wei-Sheng Lai, and Wei-Chen Chiu. Stylizing 3d scene via implicit representation and hypernetwork. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022. 3
- [8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 2021. 2
- [9] Shuangkang Fang, Yufeng Wang, Yi Yang, Yi Yang, Yi-Hsuan Tsai, Wenrui Ding, Ming-Hsuan Yang, and Shuchang Zhou. Text-driven editing of 3d scenes without retraining. *Arxiv preprint arXiv:2309.04917*, 2023. 2
- [10] William Gao, Noam Aigerman, Thibault Groueix, Vova Kim, and Rana Hanocka. Textdeformer: Geometry manipulation using text guidance. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. 3
- [11] Ayaan Haque, Matthew Tancik, Alexei Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *CVPR*, 2023. 2, 3, 6, 7, 8, 12
- [12] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2
- [13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 2
- [14] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 2022. 2
- [15] 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. 3
- [16] Hsin-Ping Huang, Hung-Yu Tseng, Saurabh Saini, Maneesh Singh, and Ming-Hsuan Yang. Learning to stylize novel views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 3
- [17] Yi-Hua Huang, Yue He, Yu-Jie Yuan, Yu-Kun Lai, and Lin Gao. Stylizednerf: consistent 3d scene stylization as stylized nerf via 2d-3d mutual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3
- [18] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)*, 2023. 2, 3, 4, 5, 12
- [19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 5
- [20] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. *Advances in Neural Information Processing Systems*, 2022. 3
- [21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 5
- [22] Yuan Li, Zhi-Hao Lin, David Forsyth, Jia-Bin Huang, and Shenlong Wang. Climatenerf: Physically-based neural rendering for extreme climate synthesis. *arXiv e-prints*, 2022. 3
- [23] Yuhan Li, Yishun Dou, Yue Shi, Yu Lei, Xuanhong Chen, Yi Zhang, Peng Zhou, and Bingbing Ni. Focaldreamer: Text-driven 3d editing via focal-fusion assembly. *arXiv preprint arXiv:2308.10608*, 2023. 3
- [24] Hao-Kang Liu, I Shen, Bing-Yu Chen, et al. Nerf-in: Free-form nerf inpainting with rgb-d priors. *arXiv preprint arXiv:2206.04901*, 2022.
- [25] Steven Liu, Xiuming Zhang, Zhoutong Zhang, Richard Zhang, Jun-Yan Zhu, and Bryan Russell. Editing conditional radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021. 3
- [26] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 5
- [27] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. *arXiv preprint arXiv:2308.09713*, 2023. 8
- [28] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3
- [29] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf:

- Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 2021. 2
- [30] Ashkan Mirzaei, Tristan Aumentado-Armstrong, Marcus A. Brubaker, Jonathan Kelly, Alex Levinstein, Konstantinos G. Derpanis, and Igor Gilitschenski. Watch your steps: Local image and scene editing by text instructions. In *arXiv preprint arXiv:2308.08947*, 2023. 2, 3
- [31] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 2022. 2
- [32] Thu Nguyen-Phuoc, Feng Liu, and Lei Xiao. Snerf: stylized neural implicit representations for 3d scenes. *arXiv preprint arXiv:2207.02363*, 2022. 3
- [33] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2
- [34] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 3
- [35] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 2019. 5
- [36] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 12
- [37] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2
- [38] 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*, 2022. 2
- [39] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 2
- [40] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, 2022. 2
- [41] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 2022. 2
- [42] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image superresolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2
- [43] Sara Fridovich-Keil and Alex Yu, Matthew Tancik, Qinzhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, 2022. 2
- [44] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2015. 2
- [45] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 2019. 2
- [46] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *CVPR*, 2022. 2
- [47] Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. Neural feature fusion fields: 3d distillation of self-supervised 2d image representations. In *2022 International Conference on 3D Vision (3DV)*, 2022. 3
- [48] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3
- [49] Can Wang, Ruixiang Jiang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Nerf-art: Text-driven neural radiance fields stylization. *TVCG*, 2023. 2, 3
- [50] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Wang Xinggang. 4d gaussian splatting for real-time dynamic scene rendering. *arXiv preprint arXiv:2310.08528*, 2023. 8
- [51] Qiling Wu, Jianchao Tan, and Kun Xu. Palettenerf: Palette-based color editing for nerfs. *arXiv preprint arXiv:2212.12871*, 2022. 3
- [52] Jiale Xu, Xintao Wang, Yan-Pei Cao, Weihao Cheng, Ying Shan, and Shenghua Gao. Instructp2p: Learning to edit 3d point clouds with text instructions. *arXiv preprint arXiv:2306.07154*, 2023. 3
- [53] Tianhan Xu and Tatsuya Harada. Deforming radiance fields with cages. In *European Conference on Computer Vision*, 2022.
- [54] Bangbang Yang, Chong Bao, Junyi Zeng, Hujun Bao, Yinda Zhang, Zhaopeng Cui, and Guofeng Zhang. Neumesh: Learning disentangled neural mesh-based implicit field for geometry and texture editing. In *European Conference on Computer Vision*, 2022. 3
- [55] Taoran Yi, Jiemin Fang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussian-dreamer: Fast generation from text to 3d gaussian splatting with point cloud priors. *arxiv:2310.08529*, 2023. 12, 13
- [56] Kai Zhang, Nick Kolkin, Sai Bi, Fujun Luan, Zexiang Xu, Eli Shechtman, and Noah Snavely. Arf: Artistic radiance fields. In *European Conference on Computer Vision*, 2022. 3

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



Figure 9. GaussianEditor demonstrates excellent extension capabilities. It can be seamlessly integrated with the 3D generative model, such as GaussianDreamer [55].

A. Appendix

A.1. Additional Implementation Details

GaussianEditor takes a 3D scene reconstructed by 3D Gaussian Splatting [18] as input. Learning each scene takes 30,000 iterations. Images wider than 512 pixels are resized to 512. Similar to Instruct NeRF2NeRF (IN2N) [11], GaussianEditor also uses Instruct Pix2Pix (IP2P) [4] to edit 2D pictures. The classifier-free diffusion guidance weights are set as follows:

- 1) Fig. 1: $s_I \in [1.4, 1.5]$, $s_T \in [7.0, 12.0]$,
- 2) Fig. 4 Bicycle: $s_I = 1.2$, $s_T = 12.0$,
- 3) Fig. 4 Bear: $s_I = 1.5$, $s_T = 6.5$,
- 4) Fig. 5: $s_I = 1.2$, $s_T = 8.0$,
- 5) Fig. 6: $s_I \in [1.2, 1.5]$, $s_T \in [7.5, 12.0]$,
- 6) Fig. 7: $s_I = 1.3$, $s_T = 12.0$,

where s_I is the weight for image guidance and s_T is the weight for text guidance.

GaussianEditor implements 3D editing based on the 2D diffusion model. Due to the instability of 2D editing, scenes tend to become blurry as the number of iterations increases. Therefore, we observe the current rendering results during the training process and limit the editing rounds, generally within 200 rounds.

A.2. Quantitative Evaluation

Quantitative Evaluation Based on CLIP. In Tab. 1, we present the quantitative evaluation results. We follow the metrics used in Instruct NeRF2NeRF (IN2N) [11], including the CLIP [36] text-image direction similarity and image-image similarity between the original scene and the edited scene. The quantitative results indicate that our method achieves a comparable CLIP text-image direction similarity score with IN2N, while image-image similarity has improved a lot. We would like to analyze the limitations of the used metric as follows.

Limitation of The CLIP-based Metric. Although we provide quantitative analysis based on CLIP. However, we

Table 1. Results of CLIP Text-Image Direction Similarity and Image-Image Similarity between the original scene and edited scene.

	CLIP Text-Image Direction Similarity ↑	Image-Image Similarity ↑
IN2N	0.12	0.86
Ours	0.11	0.94

"This is white"	0.192	0.231
"This is yellow"	0.176	0.272

Figure 10. Similarity scores between the text and image features encoded by CLIP [36]. Pure white images consistently have lower scores³.

find that the current CLIP-based metrics are not reliable enough. For example, CLIP has problems with color discrimination. As shown in Fig. 10, we use CLIP to calculate the similarity between solid color images, which are white and yellow respectively, and the text descriptions, *i.e.* “This is white” or “This is yellow”. The results show that yellow images consistently achieve higher matching scores. This is one of the reasons why our CLIP text-image direction similarity does not show an evident advantage. Therefore, we believe that a more reliable evaluation metric for text-guided editing tasks is one of the important future research directions.

A.3. Extension

GaussianEditor demonstrates excellent extension abilities. For instance, it can be seamlessly integrated with the 3D

³The red border is to make it easier for readers to see the white image. The actual image input to the CLIP does not have this border.

generative model GaussianDreamer [55], resulting in enhanced editing effects. Specifically, as shown in Fig. 9, upon obtaining the Gaussian RoI, the Gaussians within the RoI are saved individually and utilized as the initialization for the 3D-generation model. Simultaneously, the text description of the edited scene is fed into the pipeline of the 3D generation model. Eventually, the edited new object is merged into the original scene to form an edited 3D scene.