

DGE: Direct Gaussian 3D Editing by Consistent Multi-view Editing

Minghao Chen, Iro Laina, and Andrea Vedaldi

Visual Geometry Group, University of Oxford
{minghao, iro, vedaldi}@robots.ox.ac.uk
silent-chen.github.io/DGE

Abstract. We consider the problem of editing 3D objects and scenes based on open-ended language instructions. A common approach to this problem is to use a 2D image generator or editor to guide the 3D editing process, obviating the need for 3D data. However, this process is often inefficient due to the need for iterative updates of costly 3D representations, such as neural radiance fields, either through individual view edits or score distillation sampling. A major disadvantage of this approach is the slow convergence caused by aggregating inconsistent information across views, as the guidance from 2D models is not multi-view consistent. We thus introduce the *Direct Gaussian Editor* (DGE), a method that addresses these issues in two stages. First, we modify a given high-quality image editor like InstructPix2Pix to be multi-view consistent. To do so, we propose a training-free approach that integrates cues from the 3D geometry of the underlying scene. Second, given a multi-view consistent edited sequence of images, we *directly* and efficiently optimize the 3D representation, which is based on 3D Gaussian Splatting. Because it avoids incremental and iterative edits, DGE is significantly more accurate and efficient than existing approaches and offers additional benefits, such as enabling selective editing of parts of the scene.

1 Introduction

Recent breakthroughs in 2D and 3D generation [28, 56, 70, 73] have significantly advanced AI-based content creation and editing. This progress is particularly impactful for 3D content, which traditionally requires highly specialized skills and years of experience. Consequently, these innovations are unlocking new creative possibilities for artists and non-professional users alike.

In this work, we consider the problem of editing 3D models based on textual instructions. Recent progress in this area is largely due to the introduction of new radiance field representations for objects and scenes, such as NeRFs [61], which are both high-quality and flexible. Because they are differentiable, radiance fields integrate well with applications of machine learning to image generation. A notable example is InstructPix2Pix (IP2P) [4], a denoising diffusion model [73, 78] that can edit an image based on textual instructions [4]. A significant benefit of differentiable 3D representations, when combined with image generation models, is that they eliminate the need for 3D annotated data. As a result, several

works have repurposed such models to edit 3D content instead, primarily in two ways. The first was pioneered by InstructNeRF2NeRF [25], which utilizes IP2P to alternate between editing a rendered view of the 3D model and updating the latter by training on the edited images. The second approach [34, 75, 112] is to update the model in a distillation-like fashion [70]. Both approaches rely on iterative mechanisms to incorporate the edits into the 3D model, so that a single edit often requires several minutes, or even hours, to complete.

In this paper, we introduce the *Direct Gaussian Editor* (DGE), a 3D editor that addresses the shortcomings of previous attempts. We design our method with three goals in mind: (i) high fidelity, (ii) high efficiency, and (iii) selective editing (*i.e.*, editing only a specific part of the scene). To achieve these goals, we propose to change both the 3D representation and the update mechanism.

As a representation, we employ 3D Gaussian Splatting (GS) [37], a radiance field model that is notably orders of magnitude faster than NeRF and NeRF-adjacent models [7, 9, 80] for both rendering and gradient computation. Additionally, GS offers another important advantage: as an explicit 3D representation made of local 3D primitives, the Gaussians, it supports local edits easily and efficiently, as long as the relevant Gaussians can be identified. In practice, one can easily identify the Gaussians by fusing the output of a 2D segmenter [39] from several rendered views of the scene.

While GS can significantly speed up the 3D editing process, it does not remove the bottleneck caused by the need to iterate several times between rendering, reconstruction, and evaluation of the underlying image-based diffusion model. The iterative methods used in prior works are slow because of the lack of view-consistent edits that could be used to update the 3D model coherently. Since 2D editors like IP2P are *monocular* and provide a distribution over all possible edits, the probability of drawing independently two or more consistent edits from different views is nearly zero. As a result, iterative dataset updates or distillation are ways of incrementally reaching a “multi-view” consensus with respect to the application of the monocular editor to such views.

Our main contribution is a more efficient alternative to this slow iterative process. We propose a method to sample *multi-view consistent* edits such that the 3D model can be updated by *directly* fitting it to the edited views. Our approach is inspired by recent progress in video generation and editing and, in particular, methods that extend image generators to video without additional training [6, 21, 38, 71, 94]. The key insight is that multiple views of the 3D model can be interpreted as an orbital video of a static scene generated by a moving camera, making multi-view consistency analogous to temporal consistency in video editing. To achieve multi-view consistency we adopt the spatio-temporal attention mechanism used in the video editing literature and extend it with additional epipolar constraints, which we can enforce due to the 3D nature of our problem. This method ensures that edits are coherent across multiple views, improving the efficiency and effectiveness of the 3D editing process.

Through qualitative and quantitative comparisons with prior works, we demonstrate two key advantages of DGE, even compared to recent works such as the

GaussianEditor [13], which also uses Gaussian Splatting. First, as a direct editing approach, DGE results in a noticeable speed boost taking approximately 4 min for a single edit. Second, ensuring multi-view consistent editing in the image space significantly simplifies the process of consolidating edits from various views into the 3D model. This is reflected in both the number of updates required for convergence and the higher fidelity of the final result.

2 Related Work

Image Editing. Due to the lack of large-scale training data to learn 3D editors directly, most such editors build on top of existing 2D image editors instead, so we discuss those first. GLIDE [66] controls 2D image generation and editing using CLIP features [72]. Methods like [19, 42] and DreamBooth [74] consider the problem of personalizing an image; GLIGEN [50] and others [8, 11, 17] perform layout control, while ControlNet [106] considers additional forms of control such as scribbles or depth maps. DragDiffusion [76] and Drag-a-Video [82] edit images and videos by dragging. Others [2, 27, 36, 57, 64, 68, 85] cast image editing as image-to-image translation. InstructPix2Pix (IP2P) [4] fine-tunes Stable Diffusion [73] with image-and-text conditioning on a large synthetic dataset of language-driven edits. The works of [105, 108] further improve InstructPix2Pix via manual labeling, while [62] aims at improving the localization of the edits.

Ad-hoc 3D Editing. Several authors have explored various types of inputs and controls to edit 3D objects. EditNeRF [53] updates shape and color in a radiance field based on user-provided scribbles, 3Designer [47] and SINE [1] based on a single edited view, SKED [59] based on 2D sketches, Editable-NeRF [109] based on keypoints and CoNeRF [35] based on attributes. The work of [53] also considers sketch-based editing but for category-level 3D generators. NeRF-Editing [101], NeuMesh [98] and NeRFShop [33] modify a radiance field based on meshes, NeuralEditor [10] based on point clouds and [97] based on cages. N3F [84], DFF [40], SPIIn-NeRF [63], NeRF-in [52] and the work of [93] consider segmenting and removing objects from a radiance field. Control-NeRF [43] allows both removing and moving objects. Component-NeRF [51] and the works of [99, 103] consider compositional editing. Palette-NeRF [41], RecolorNeRF [22], ICE-NeRF [44] address recoloring and ARF [104], DeSRF [96], StylizedNeRF [31], SNeRF [65] and the work of [15, 30] style transfer. NeRFEditor [81] integrates GAN-based stylization in NeRF. Seal-3D [92] edits 3D scenes by learning interactive tools like brushes, deformation, and recoloring and Seal4D-NeRF [32] extends it to dynamic NeRF models. SceNeRFFlow [83] allows for dynamic edits using DensePose for correspondence estimation [24].

Language-driven 3D Editing. Closer to our work, several authors have considered open-vocabulary or text-guided 3D editing. Some authors use vision-language models based on CLIP [46, 48, 72] to edit or stylize a radiance field

globally [45, 58, 88, 89] or locally [23, 40, 77, 90, 110]. TextDeformer [20] focuses on manipulating only the shape of objects. Some like AvatarCLIP [29] consider animations as well. Most recent works have shifted towards instruction-guided editing employing diffusion-based image generators like Stable Diffusion [73] or editors like Instruct-Pix2Pix [4]. There exist two main editing mechanisms among them. Instruct-NeRF2NeRF [25], InstructP2P [95], ProteusNeRF [87], Edit-DiffNeRF [100], and GaussianEditor [13, 91] repeatedly edit rendered views of the 3D object and update the 3D model accordingly, a process which is referred to as iterative dataset updates. Instead, Instruct3Dto3D [34], DreamEditor [112], Vox-E [75], ED-NeRF [67], FocalDreamer [49], Progressive3D [14] and [102, 111] update the 3D model using score distillation sampling (SDS). While these works rely on costly, scene-specific optimization of radiance fields, SHAP-EDITOR [12] learns a fast, feed-forward editor in the latent space of generative models but requires retraining for each new set of instructions. Our editor is optimization-based, but significantly reduces the number of iterations required, up to a single one.

Closely related to our work, ViCa-NeRF [16] focuses on multi-view consistent editing. However, different from our method, which takes inspiration from video models, ViCA-NeRF leverages the 3D model’s depth information to project features from key views into others using a blending module. As a result, it cannot handle edits that change the geometry of the original 3D shape.

Furthermore, methods like Vox-E [75] and FocalDreamer [49] consider local 3D editing, but struggle to handle both local and global ones; in contrast, our DGE allows precise editing of small or large regions of the 3D scene.

3 Preliminaries

Gaussian Radiance Fields. A *radiance field* is a pair of functions $\sigma : \mathbb{R}^3 \rightarrow \mathbb{R}^+$ and $c : \mathbb{R}^3 \times \mathbb{S}^2 \rightarrow \mathbb{R}^3$ mapping 3D points $\mathbf{x} \in \mathbb{R}^3$ to opacities $\sigma(\mathbf{x})$ and directional colors $c(\mathbf{x}, \boldsymbol{\nu})$, where $\boldsymbol{\nu} \in \mathbb{S}^2$ is a unit vector expressing the viewing direction. An image I of the radiance field is obtained via the *emission-absorption* equation:

$$I(\mathbf{u}) = \int_0^\infty c(\mathbf{x}_t, \boldsymbol{\nu}) \sigma(\mathbf{x}_t) e^{-\int_0^t \sigma(\mathbf{x}_\tau) d\tau} dt, \quad (1)$$

where the 3D point $\mathbf{x}_t = \mathbf{x}_0 - t\boldsymbol{\nu}$ sweeps the ray that connects the camera center \mathbf{x}_0 to the pixel \mathbf{u} along direction $-\boldsymbol{\nu}$ (where the 3D point is expressed in the reference frame of the world, not the camera).

Prior works have explored several representations for these functions, including MLPs [61], voxel grids [80] and low-rank factorizations of the latter [7, 9]. *Gaussian Splatting* (GS) [37] proposes a particularly efficient representation as a mixture of Gaussians $\mathcal{G} = \{(\sigma_i, \boldsymbol{\mu}_i, \Sigma_i, c_i)\}_{i=1}^C$, where $\sigma_i \geq 0$ is the opacity, $\boldsymbol{\mu}_i \in \mathbb{R}^3$ is the mean, $\Sigma_i \in \mathbb{R}^{3 \times 3}$ is the covariance matrix, and $c_i : \mathbb{S} \rightarrow \mathbb{R}^3$ the directional color of each Gaussian. The Gaussian functions are given by:

$$g_i(\mathbf{x}) = \exp \left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)^\top \Sigma_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \right). \quad (2)$$

The directional colors are given by functions $[c_i(\boldsymbol{\nu})]_j = \sum_{l=0}^L \sum_{m=-l}^l c_{ijlm} Y_{lm}(\boldsymbol{\nu})$, where Y_{lm} are spherical harmonics and $c_{ijlm} \in \mathbb{R}$ are corresponding coefficients. The Gaussian mixture then defines the opacity and color functions as:

$$\sigma(\mathbf{x}) = \sum_{i=1}^G \sigma_i g_i(\mathbf{x}), \quad c(\mathbf{x}, \boldsymbol{\nu}) = \frac{\sum_{i=1}^G c_i(\boldsymbol{\nu}) \sigma_i g_i(\mathbf{x})}{\sum_{i=1}^G \sigma_i g_i(\mathbf{x})}. \quad (3)$$

Importantly, the integral in Eq. (1) under the model in Eq. (3) can be approximated very efficiently and in a differentiable manner, as noted in [37].

Diffusion-based Generators and Editors. A *diffusion-based image generator* is a model that allows one to sample image $I \in \mathbb{R}^{3 \times H \times W}$ from a conditional distribution $p(I | y)$. Here the condition y is often a textual prompt, but it can also be an image or a combination of both. Let $I_k = \sqrt{1 - \beta_k^2} I + \beta_k \epsilon$ be a noised version of the image I one wishes to sample, where β is a sequence monotonically increasing from 0 to 1 and $\epsilon \sim \mathcal{N}$ is normal i.i.d. noise. The diffusion model consists of a denoising neural network Φ that, given the noisy image I_k , the noise level k and the conditioning y , estimates the noise $\hat{\epsilon} = \Phi(I, y, k)$. The image I is sampled by starting from pure noise ($\beta_T = 1$) followed by iterative denoising of the signal until $I = I_0$ is obtained [28, 78, 86].

Several authors [27, 57, 68] have suggested using the denoising network of an off-the-shelf image generator such as Stable Diffusion (SD) [69, 73] to modify an existing image according to a prompt y instead of generating it from scratch, thus editing it. However, this process is somewhat suboptimal; instead, we make use of editors specifically trained to apply such modifications. An example is IP2P [4], which also uses diffusion but directly implements conditional a distribution $p(I' | I, y)$ where I' is the edited image, I is the original image, and y is a prompt describing how I' should be obtained from I . IP2P is in itself derived from the SD model by finetuning it to training triplets (I', I, y) which are automatically generated using Prompt-to-Prompt [27], a slow but training-free editing method.

4 Method

In this section, we describe our method for text-guided 3D scene editing—the Direct Gaussian Editor (DGE). Our main objective is 3D editing that is (i) faithful to the text instructions (high fidelity), (ii) fast and efficient, and (iii) partial to specific scene elements when desired. Next, we introduce our approach and discuss how these criteria are met.

4.1 Direct Gaussian Editor

DGE is motivated by the following key observation: given several images rendered from an existing 3D model, if one can obtain *multi-view consistent edits* in image space, then the 3D model can be updated *directly* and efficiently by

fitting it to the edited images. This approach is an alternative to slow iterative techniques like SDS [49, 70, 75] and iterative dataset updates (IDU) [13, 18, 25] that work around inconsistent image-level edits. Thus, the main challenge for enabling direct 3D editing is to improve image-based editors to produce several *consistent* edits across a number of images.

In more detail, consider an initial reconstruction of a 3D scene as a mixture of Gaussians \mathcal{G} (Sec. 3) and consider a camera trajectory $(\pi_t)_{t=1}^T$ consisting of T viewpoints. We first render the corresponding views $\mathcal{I} = (I_t)_{t=1}^T$ where $I_t = \text{Rend}(\mathcal{G}, \pi_t)$ is obtained with Eq. (1). Let y be the textual prompt describing the desired edits and let $\mathcal{I}' = (I'_t)_{t=1}^T$ be the images obtained by editing \mathcal{I} as according to y . We can then update the 3D model \mathcal{G} by fitting it to \mathcal{I}' as

$$\mathcal{G}' = \arg \min_{\mathcal{G}} \sum_{t=1}^T \|I'_t - \text{Rend}(\mathcal{G}, \pi_t)\| \quad (4)$$

This rendering loss is often used to reconstruct a 3D scene from multiple views, and here it is used to update it. In practice, GS makes this optimization efficient.

The key question is how to obtain the edited views \mathcal{I}' . A standard approach is to apply a diffusion-based image editor \mathcal{E} , such as IP2P [95] to *individual* views, obtaining edited views \mathcal{I}' . However, each of these views would then be drawn *independently* from the distribution $I'_t \sim p(I'_t | I_t, y)$. Hence, even though the initial views I_t are consistent by construction, the edited views would not be as this method disregards their statistical dependency. As a result, updating the 3D model \mathcal{G} according to Eq. (4) would fuse inconsistent views, yielding a blurry outcome. We thus seek to modify the image editor to account for the mutual dependency of the views and approximately draw samples from the joint distribution $p(I'_1, \dots, I'_T | I_1, \dots, I_T, y)$.

In summary, as shown in Figure 1, we divide the 3D editing process into two parts, detailed next: (a) multi-view consistent editing with epipolar constraints, and (b) the 3D reconstruction from edited images.

4.2 Consistent Multi-view Editing with Epipolar Constraint

To achieve multi-view consistent edits, we render multiple views \mathcal{I} from \mathcal{G} and reorder them based on their camera positions to create a relatively smooth camera trajectory, forming a *video*. This approach leverages recent advances in video generation and editing, and in particular, the idea of extending image generators to video without additional training [6, 21, 38, 71, 94]. Treating the sequence \mathcal{I} as a video of a static scene and applying these techniques to image *editors*, consistency across views is analogous to temporal consistency in video editing.

Our multi-view editing process includes two sequential steps, key-view editing and feature injection. Specifically, during each denoising step, we first jointly edit select key views with *spatio-temporal* attention. Then, we inject the feature of the key views to other views using correspondences that are obtained from visual features guided by the epipolar constraint.

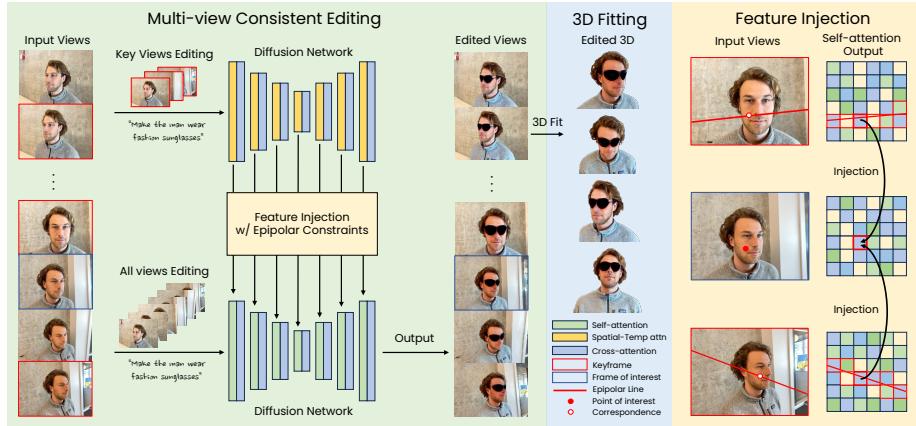


Fig. 1: Overview. As shown on the left, our method is divided into two main parts: multi-view consistent editing with epipolar constraints and direct 3D fitting. In the multi-view editing stage, key views are randomly selected and jointly fed to the editing diffusion network to extract features with the *spatial-temporal attention*. To edit other frames, the features of key views are injected into the diffusion network through correspondence matching on feature maps with epipolar constraints. The detailed feature injection process is shown on the right; only features with a red border (*i.e.*, the points following epipolar constraints) are considered for correspondence matching.

Key-view editing. Let \mathcal{E} be a diffusion-based image editor that utilizes self-attention (*e.g.*, SD-based). Based on the above observations, we extend \mathcal{E} with a modified, *spatio-temporal* version of self-attention, which jointly considers all images in \mathcal{I} , effectively turning the single-image editor into a multi-image editor \mathcal{E}_{mv} . Specifically, given queries $\{Q_t\}_{t=1}^T$, keys $\{K_t\}_{t=1}^T$, and values $\{V_t\}_{t=1}^T$ for each viewpoint t , the spatio-temporal attention block ensures that each frame attends to every other frame and is given by:

$$\text{STAttn}(Q, K, t) = \text{Softmax}\left(\frac{Q_t \cdot [K_1, \dots, K_T]}{\sqrt{d}}\right), \quad (5)$$

where d is the embedding dimension of keys and queries. The corresponding output features for frame t are then computed as $\Phi_t = \text{STAttn}(Q, K, t) \cdot [V_1, \dots, V_T]$ (at each layer). As the underlying image editor \mathcal{E} , we utilize IP2P [4], which is a common choice in the literature [13, 25, 91]. However, the attention can be computationally expensive for longer sequences. Therefore, \mathcal{E}_{mv} is only applied to select key views randomly selected from $\mathcal{T} = \{1, \dots, T\}$, where T is the total number of images in the sequence. In practice, we sample a random subset $\mathcal{K} \subset \mathcal{T}$ at each denoising iteration. We detail the selection process in the supplement. This step results in roughly consistent edits across all key views.

Feature injection. The goal of this step is to propagate the edited *key view* features to *all other* views to ensure the consistency of the edits. Specifically,

we inject the features by finding inter-frame correspondences with epipolar constraints. To find the correspondences between key views and other views, we first extract visual features Ψ_t for all views in different layers of the denoising network Φ , using the intermediate outputs of Φ (*i.e.*, the inputs to each self-attention block). Then, point correspondences across views can be established by simply comparing their respective features at all spatial locations, and the key view features can be injected (based on these correspondences) into all remaining views, avoiding expensive self-attention blocks.

In addition, we can also leverage the fact that we can directly use 3D information to constrain the correspondence problem. In fact, we have multiple views of the scene available, and the pose and calibration matrix of cameras are known. This is without loss of generality since these assumptions are required to apply GS for scene reconstruction in the first place. Therefore, we can estimate the fundamental matrix F between two views and constrain the correspondence problem to points along an epipolar line.

Formally, given a feature map $\Psi_{t'}$ corresponding to image $I_{t'}$, where $t' \notin \mathcal{K}$, and the features Ψ_{k^*} of a keyframe I_{k^*} , the correspondence map $M_{t'}$ is given by:

$$M_{t'}[u] = \arg \min_{v, v^\top F u = 0} D(\Psi_{t'}[u], \Psi_{k^*}[v]), \quad \forall t' \in \mathcal{T} \setminus \mathcal{K} \quad (6)$$

where D is the cosine distance, u and v index the feature maps spatially, k^* is the index of the *key* view that is the closest to view t' (in terms of camera viewpoint), and F is the fundamental matrix corresponding to the two views t' and k^* . Fu is the epipolar line in view k^* along which the corresponding point of v in view k^* must lie [26]. Intuitively, using the epipolar constraint significantly reduces the search space of correspondences from a plane to spatial locations along a single line. Given this constraint, the correspondence is decided based on the cosine distance of the features. In practice, we compute the correspondence maps between a view t' and its two nearest neighbors. For a given position u , we linearly combine the features of correspondences in the nearest two key views to obtain its edited feature.

This approach is similar in spirit to TokenFlow [21] but contributes two noteworthy improvements that are tailored to the 3D editing task in particular. First, using IP2P instead of the 2D editor [85] used in TokenFlow avoids the time-consuming DDIM [79] inversion process. Second, the epipolar constraint in Eq. (6) allows us to leverage the 3D information that is available from the original model \mathcal{G} when propagating edits. This, in turn, can be particularly useful in cases where appearance is insufficient to compute correspondences (*e.g.*, mostly uniform appearance, as shown in Fig. 4).

4.3 Direct Reconstruction and Iterative Refinement

After obtaining a set of edited images \mathcal{I} with roughly consistent content, we can directly fit the Gaussian mixture \mathcal{G} to the resulting images using Eq. (4), without the need for distillation (SDS). One of the notable advantages of 3D

GS over NeRF is that it enables the rendering of images at full resolution and the application of robust image-level losses. Specifically, we optimize \mathcal{G} using LPIPS [107] as an objective function between the edited I'_t and the rendered \hat{I}_t views in Eq. (4). The LPIPS loss is insensitive to small inconsistencies in the edited images, and helps to obtain a more consistent 3D output.

Finally, we observe that even though \mathcal{E}_{mv} results in generally consistent edits, some inconsistencies and artifacts may remain, especially around fine and detailed textures, which in turn lead to blurry results after updating \mathcal{G} . We conjecture that this is due to the limited resolution of the feature space. To address this issue, after obtaining \mathcal{G}' , we can again render images to be edited, repeating the process described in the previous section and re-updating the 3D model, similar to SDEdit [57] or IM-3D [55].

4.4 Partial Editing

Because 3D GS is an *explicit* 3D representation, if desired, a user can selectively edit a 3D scene by allowing only certain Gaussians to change, thus focusing on a specific region of interest while ensuring that the rest of the scene remains unchanged. This is a significant advantage compared to implicit radiance field representations, where there is no direct correspondence between specific parameters and regions of space, so any update to the model tends to be global.

We thus propose to first obtain a mask for the Gaussians that should undergo editing. This could be done following existing works [5, 13]. In this paper, we mainly adopt the approach in [13], which segments the 2D views first and unprojects the 2D segmentation results to 3D to obtain the masks of Gaussians.

We then follow the same steps outlined in Secs. 4.2 and 4.3, but only optimize the reconstruction objective on regions defined by the rendered segmentation masks. This training procedure leads to the partial change of a 3D model when a specific prompt is provided.

5 Experiments

In this section, we first provide the implementation details of our method followed by qualitative and quantitative comparison with other methods. We then provide an ablation study of our method.

Implementation Details. As our image editor, we use InstructPix2Pix [4] for a fair comparison to prior methods using the same model. We use scenes from IN2N [25] and other real-world datasets, including LLFF and Mip-NeRF360 [3, 60] to demonstrate the ability of our method to edit 3D models. We use 3D GS [37] as the 3D representation. For editing, we use 20-30 views, which are edited according to the approach outlined in Sec. 4.2 and then train \mathcal{G}' with 500-1500 iterations depending on the complexity of the scenes. We use LPIPS and L1 loss to train the 3D GS as suggested in IN2N [25]. We also use classifier-free guidance to control the effect strength of the editing. Most of the edits use

Method	3D	CLIP	CLIP Directional	Avg. Editing
	Model	Similarity	Similarity	Time
Instruct-N2N [25]	NeRF	0.215	0.64	~ 51min
ViCA-NeRF [16]	NeRF	0.204	0.44	~ 28min
GaussianEditor [13]	GS	0.201	0.60	~ 7min
IP2P [95] + SDS [70]	GS	0.206	0.61	~ 6min
Ours	GS	0.226	0.67	~ 4min

Table 1: Comparison with other editing methods. Methods based on Gaussian Splatting are much faster than NeRF-based ones. Our DGE achieves the best performance at almost half the time compared to GaussianEditor.

the default setting, 7.5 for textual conditions and 1.5 for the image condition. We perform the iterative refinement every 500 iterations. We provide more details in the supplementary material.

Evaluation. We provide both qualitative and quantitative evaluations of our method. For quantitative evaluations, we follow common practice [13, 25]. We evaluate the alignment of edited 3D models and target text prompts with CLIP similarity score, *i.e.*, the cosine similarity between the text and image embeddings encoded by CLIP. Specifically, we randomly sample 20 camera poses from the training dataset of the 3D models and measure the CLIP similarity between the rendered images and the target text prompt. Additionally, to measure the editing effect, we use the CLIP directional similarity, *i.e.*, the cosine similarity between the image and text editing directions (target embeddings minus source embeddings). We evaluate all methods on 3 different scenes and 10 different prompts. The detailed scene-prompt pairs are provided in the supplement.

5.1 Comparisons with Prior Work

We first compare our method with representative text-guided 3D editing methods, IN2N [25] and GaussianEditor [13]. They both employ the iterative dataset updates (IDU) approach, while the former builds on a NeRF representation and the latter on Gaussian Splatting. To edit a 3D model, they both iterate between the 3D model training and image editing. Every 10 steps, they use the same 2D editor (IP2P) to edit the rendered image, replace the edited image in the 3D model training dataset, and continue the model training. In addition, we include a baseline using the same GS framework as in our approach, but applying edits via SDS with InstructPix2Pixel (IP2P + SDS). This baseline enables a direct and fair comparison between our consistent multi-view editing approach and the popular SDS-based approach. Lastly, we compare with ViCA-NeRF [16], which also focuses on editing different views consistently. Different from ViCA-NeRF, which uses 3D model depth information to project features from key views into

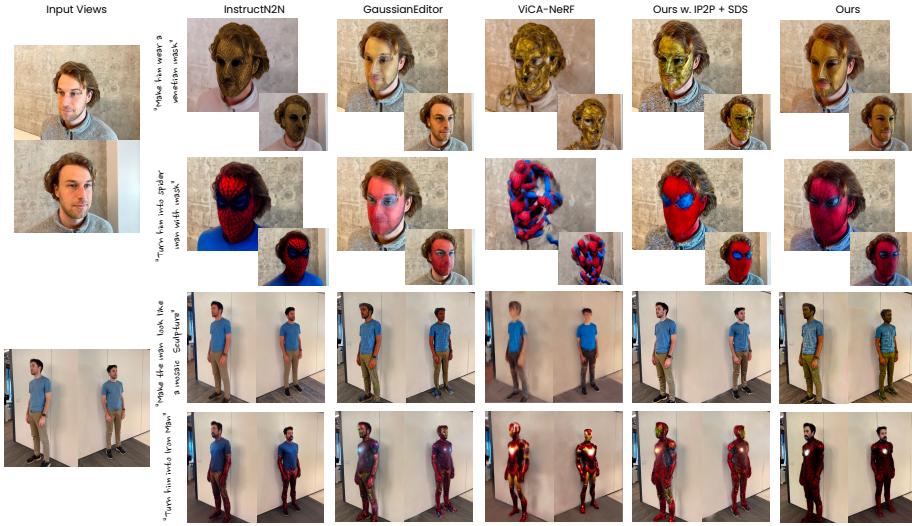


Fig. 2: Comparison with other methods. Our method can provide fast and detailed editing effects, such as the textures on the Venetian mask and mosaic sculpture. Other methods, such as InstructN2N and IP2P+SD, fail to get the mosaic effects because they average over inconsistent editing.

Multi-view Editing	Epipolar Constraints	CLIP Score	CLIP Directional Score
		0.203	0.59
	✓	0.221	0.64
✓	✓	0.226	0.67

Table 2: Ablation study on different components of our methods. Without multi-view consistency editing, both the CLIP and CLIP directional scores drop significantly. Although the difference between with and without epipolar constraints is small, we provide a visual comparison in Fig. 4 showing that it helps to improve the detailed texture especially when the appearance is relatively similar in different places.

others using a blending module, our approach incorporates the constraints from epipolar lines to avoid depth estimation errors.

In Table 1 we compare DGE to the above methods and baselines in terms of CLIP similarity score and CLIP directional similarity score. Our approach outperforms all prior work while cutting the editing time to almost half compared to GaussianEditor, the second fastest alternative.

In Figure 2 we demonstrate visual comparisons with the above methods. Our proposed method achieves realistic and detailed 3D edits. One notable advantage of our approach is that ensuring the consistency of the edits in image space results in 3D edits that are overall more faithful to the editing instruction. In

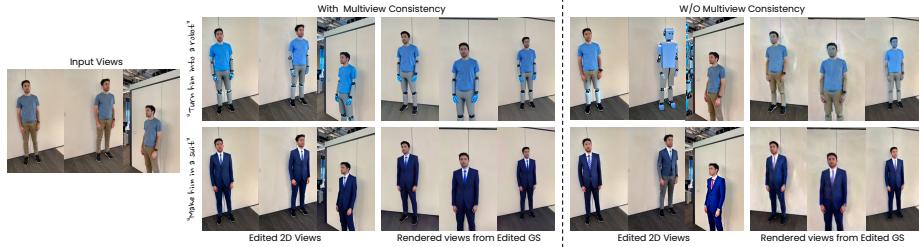


Fig. 3: The comparison between with and without multi-view consistency. With the proposed multi-view consistent editing, the edited 3D GS is clear and clean, while without it, it either fails to converge or leads to blurry results.



Fig. 4: The comparison between edited 2D images with and without epipolar constraints. The one with epipolar constraints successfully matches the correspondences, while the other fails, thus resulting in inconsistent multi-view edits.

contrast, IDU and SDS work by progressively aligning the edited views of the 3D object, averaging the inconsistencies from the edited images and leading to blurrier or lower-fidelity reconstructions and artifacts. For example, for the edit instruction “*Make him wear a Venetian mask*”, the texture of the mask generated by DGE is detailed and vivid, while other methods fail to properly generate the texture. ViCA-NeRF results in blurry edits or even fails to produce meaningful edits given prompts such as “*Turn him into spider man with a mask*”. It is worth noting that IN2N and ViCA-NeRF generally yield smoother results than GaussianEditor and IP2P+SDS; this is due to the fact the NeRF is continuous, while GS is not.

5.2 Ablation Study

Next, we ablate the effectiveness of two main components in our editing pipeline: the multi-view consistent editing and the epipolar correspondence matching. Finally, we demonstrate the convergence speed (in terms of editing) of our approach through a qualitative comparison with [13].

Editing without Multi-View Consistency. To demonstrate the effectiveness of the multi-view consistent editing, we replace the multi-view consistent editing, by *independently* editing the same views and use those to fit \mathcal{G}' . Table 2 shows the importance of multi-view consistency edits, achieving higher CLIP and CLIP directional scores. To further demonstrate this, in Fig. 3 we show some



Fig. 5: Comparison between our DGE and GaussianEditor [13] in terms of the number of iterations. Our method achieves realistic editing results with much fewer iterations. With more iterations, our method also gradually refines the details.

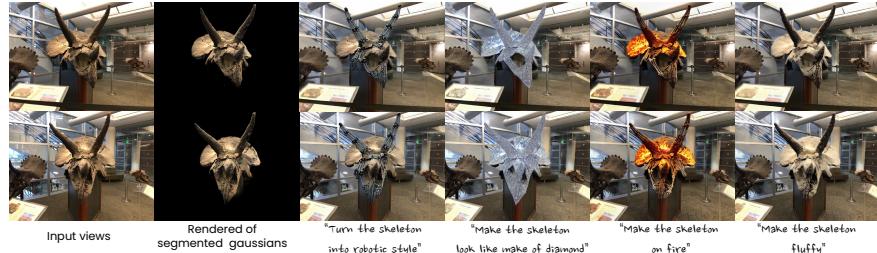


Fig. 6: Partial editing results on the *Horns* scene. We achieve realistic editing results only on the required object while keeping the rest unchanged.

intermediate 2D editing results with and without multi-view consistent editing. We observe that multi-view consistent editing produces views with similar appearances, but views can be quite different without it. This is especially true in the first example (“*turn the man into a robot*”), where the prompt allows for greater variation among views and may even fail to produce sensible edits for some of the frames (the third in this example). As expected, fitting \mathcal{G}' to these edited views produces a low-quality 3D edited model. Our multi-view consistent approach clearly alleviates these issues. We show more detailed comparisons in supplementary material.

Effect of the Epipolar Constraints. In Table 2, we show that using the epipolar constraints during feature injection (*i.e.*, from key views to other views), both the CLIP score and CLIP directional score increase. Quantitatively, the impact of using the epipolar constraint appears to be relatively small. However, we

conjecture that this is mainly because the CLIP score assesses the image at a high level while the epipolar constraints help to improve the finer details in editing, such as textures. The importance of the epipolar constraint is more evident qualitatively. In Figure 4, the resulting edits without the epipolar constraint are inconsistent between the two views—red points (just below the bear’s ear). The reason is that it wrongly matches the point with another point in black which leads to failed consistent editing. The red line in the images indicates the epipolar line corresponding to the red point on the key view.

Effect of the Number of Iterations. We additionally show a comparison between our method and GaussianEditor in Figure 5. The figure demonstrates that our method achieves high-quality edits after only 400 iterations, while GaussianEditor requires much longer for convergence. In particular, the results of the 400-iterations column do not use refinement and are generated in *under two minutes*. With more iterations, the details of the mask in the rendered views are gradually refined.

5.3 Partial Editing

Finally, in Figure 6, we showcase partial (local) editing results on the *Horns* scene from the LLFF [60] dataset. First, we segment 2D renders to get 2D masks of the skeleton we want to edit. Then, we unproject the 2D masks to 3D getting the mask of the Gaussians. As last, we render images from different views and edit those views with our proposed DGE. The results are realistic and only applied to the segmented skeleton, demonstrating our method’s ability of partial editing. We provide additional partial editing results in the supplementary material.

6 Conclusion

We have presented DGE, a robust framework for *directly* editing of a 3D model by simply fitting it to a small set of views, which are edited by a text-based image editor. The key insight of our approach is that one must obtain multi-view *consistent* edits in image space, to be able to update the 3D model directly. Thus we propose an editing mechanism that jointly considers multiple frames based on both appearance cues (image features) and 3D cues (epipolar constraints based on scene geometry). This approach enables 3D editing that is both more efficient and more faithful to the text instruction in comparison to SDS-like alternatives.

Ethics. For further details on ethics, data protection, and copyright please see <https://www.robots.ox.ac.uk/~vedaldi/research/union/ethics.html>.

Acknowledgements. This research is supported by ERC-CoG UNION 101001212. I. L. is also partially supported by the VisualAI EPSRC grant (EP/T028572/1).

References

1. Bao, C., Zhang, Y., Yang, B., Fan, T., Yang, Z., Bao, H., Zhang, G., Cui, Z.: SINE: semantic-driven image-based NeRF editing with prior-guided editing field. In: CVPR (2023) [3](#)
2. Bar-Tal, O., Ofri-Amar, D., Fridman, R., Kasten, Y., Dekel, T.: Text2live: Text-driven layered image and video editing. In: ECCV (2022) [3](#)
3. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-NeRF 360: Unbounded anti-aliased neural radiance fields. In: CVPR (2022) [9](#), [20](#), [21](#), [26](#)
4. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: CVPR (2023) [1](#), [3](#), [4](#), [5](#), [7](#), [9](#), [26](#), [27](#)
5. Cen, J., Fang, J., Yang, C., Xie, L., Zhang, X., Shen, W., Tian, Q.: Segment any 3d gaussians. arXiv preprint arXiv:2312.00860 (2023) [9](#)
6. Ceylan, D., Huang, C.H.P., Mitra, N.J.: Pix2video: Video editing using image diffusion. In: ICCV. pp. 23206–23217 (2023) [2](#), [6](#)
7. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., Mello, S.D., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., Karras, T., Wetzstein, G.: Efficient geometry-aware 3D generative adversarial networks. In: CVPR (2022) [2](#), [4](#)
8. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. In: SIGGRAPH (2023) [3](#)
9. Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H.: TensoRF: Tensorial radiance fields (2022) [2](#), [4](#)
10. Chen, J., Lyu, J., Wang, Y.: NeuralEditor: Editing neural radiance fields via manipulating point clouds. In: CVPR (2023) [3](#)
11. Chen, M., Laina, I., Vedaldi, A.: Training-free layout control with cross-attention guidance. In: WACV (2023) [3](#)
12. Chen, M., Xie, J., Laina, I., Vedaldi, A.: SHAP-EDITOR: Instruction-guided latent 3D editing in seconds. In: CVPR (2024) [4](#)
13. Chen, Y., Chen, Z., Zhang, C., Wang, F., Yang, X., Wang, Y., Cai, Z., Yang, L., Liu, H., Lin, G.: Gaussianeditor: Swift and controllable 3d editing with gaussian splatting. CVPR (2024) [3](#), [4](#), [6](#), [7](#), [9](#), [10](#), [12](#), [13](#), [23](#), [24](#)
14. Cheng, X., Yang, T., Wang, J., Li, Y., Zhang, L., Zhang, J., Yuan, L.: Progressive3d: Progressively local editing for text-to-3d content creation with complex semantic prompts. In: ICLR (2024) [4](#)
15. Chiang, P.Z., Tsai, M.S., Tseng, H.Y., sheng Lai, W., Chiu, W.C.: Stylizing 3D scene via implicit representation and hypernetwork **2105.13016** (2022) [3](#)
16. Dong, J., Wang, Y.: ViCA-NeRF: View-consistency-aware 3d editing of neural radiance fields. In: NeurIPS (2024) [4](#), [10](#)
17. Epstein, D., Jabri, A., Poole, B., Efros, A.A., Holynski, A.: Diffusion self-guidance for controllable image generation. In: NeurIPS (2023) [3](#)
18. Fang, J., Wang, J., Zhang, X., Xie, L., Tian, Q.: Gaussianeditor: Editing 3d gaussians delicately with text instructions. In: CVPR (2024) [6](#)
19. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: ICLR (2023) [3](#)
20. Gao, W., Aigerman, N., Groueix, T., Kim, V., Hanocka, R.: TextDeformer: Geometry manipulation using text guidance. In: SIGGRAPH (2023) [4](#)

21. Geyer, M., Bar-Tal, O., Bagon, S., Dekel, T.: Tokenflow: Consistent diffusion features for consistent video editing. In: ICLR (2024) [2](#), [6](#), [8](#)
22. Gong, B., Wang, Y., Han, X., Dou, Q.: Recolornerf: Layer decomposed radiance field for efficient color editing of 3d scenes. In: ACM MM (2023) [3](#)
23. Gordon, O., Avrahami, O., Lischinski, D.: Blended-nerf: Zero-shot object generation and blending in existing neural radiance fields. ICCV (2023) [4](#)
24. Güler, R.A., Neverova, N., Kokkinos, I.: DensePose: Dense human pose estimation in the wild. In: CVPR (2018) [3](#)
25. Haque, A., Tancik, M., Efros, A.A., Holynski, A., Kanazawa, A.: Instruct-NeRF2NeRF: Editing 3D scenes with instructions. In: ICCV (2023) [2](#), [4](#), [6](#), [7](#), [9](#), [10](#), [23](#), [24](#), [26](#), [27](#)
26. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision (2000) [8](#)
27. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-or, D.: Prompt-to-prompt image editing with cross-attention control. In: ICLR (2023) [3](#), [5](#)
28. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS **33**, 6840–6851 (2020) [1](#), [5](#)
29. Hong, F., Zhang, M., Pan, L., Cai, Z., Yang, L., Liu, Z.: AvatarCLIP: zero-shot text-driven generation and animation of 3D avatars. SIGGRAPH **41**(4) (2022) [4](#)
30. Huang, H., Tseng, H., Saini, S., Singh, M., Yang, M.: Learning to stylize novel views. In: ICCV (2021) [3](#)
31. Huang, Y., He, Y., Yuan, Y., Lai, Y., Gao, L.: StylizedNeRF: Consistent 3D scene stylization as stylized NeRF via 2D-3D mutual learning. In: CVPR (2022) [3](#)
32. Huang, Z., Shi, Y., Bruce, N., Gong, M.: SealD-NeRF: Interactive pixel-level editing for dynamic scenes by neural radiance fields **2402.13510** (2024) [3](#)
33. Jambon, C., Kerbl, B., Kopanas, G., Diolatzis, S., Leimkühler, T., Drettakis, G.: NeRFshop: interactive editing of neural radiance fields. Proc. ACM Comput. Graph. Interact. Tech. **6**(1) (2023) [3](#)
34. Kamata, H., Sakuma, Y., Hayakawa, A., Ishii, M., Narihira, T.: Instruct 3d-to-3d: Text instruction guided 3d-to-3d conversion. arXiv preprint arXiv:2303.15780 (2023) [2](#), [4](#)
35. Kania, K., Yi, K.M., Kowalski, M., Trzcinski, T., Tagliasacchi, A.: Conerf: Controllable neural radiance fields. In: CVPR. pp. 18623–18632 (2022) [3](#)
36. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mossner, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: CVPR (2023) [3](#)
37. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3D Gaussian Splatting for real-time radiance field rendering. SIGGRAPH **42**(4) (2023) [2](#), [4](#), [5](#), [9](#)
38. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In: ICCV (2023) [2](#), [6](#)
39. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. In: CVPR (2023) [2](#)
40. Kobayashi, S., Matsumoto, E., Sitzmann, V.: Decomposing nerf for editing via feature field distillation. NeurIPS **35**, 23311–23330 (2022) [3](#), [4](#)
41. Kuang, Z., Luan, F., Bi, S., Shu, Z., Wetzstein, G., Sunkavalli, K.: PaletteNeRF: Palette-based appearance editing of neural radiance fields. In: CVPR (2023) [3](#)
42. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: CVPR (2023) [3](#)

43. Lazova, V., Guzov, V., Olszewski, K., Tulyakov, S., Pons-Moll, G.: Control-NeRF: Editable feature volumes for scene rendering and manipulation. In: WACV (2023) 3
44. Lee, J.H., Kim, D.S.: Ice-nerf: Interactive color editing of nerfs via decomposition-aware weight optimization. In: ICCV. pp. 3491–3501 (2023) 3
45. Lei, J., Zhang, Y., Jia, K., et al.: Tango: Text-driven photorealistic and robust 3d stylization via lighting decomposition. NeurIPS **35**, 30923–30936 (2022) 4
46. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. In: ICLR (2022) 3
47. Li, G., Zheng, H., Wang, C., Li, C., Zheng, C., Tao, D.: 3DDesigner: Towards photorealistic 3d object generation and editing with text-guided diffusion models [abs/2211.14108](#) (2022) 3
48. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: CVPR. pp. 10965–10975 (2022) 3
49. Li, Y., Dou, Y., Shi, Y., Lei, Y., Chen, X., Zhang, Y., Zhou, P., Ni, B.: Focal-dreamer: Text-driven 3d editing via focal-fusion assembly. In: AAAI (2024) 4, 6
50. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: CVPR (2023) 3
51. Lin, Y., Bai, H., Li, S., Lu, H., Lin, X., Xiong, H., Wang, L.: CompoNeRF: Text-guided multi-object compositional nerf with editable 3d scene layout [abs/2303.13843](#) (2023) 3
52. Liu, H., Shen, I., Chen, B.: NeRF-In: Free-form NeRF inpainting with RGB-D priors [abs/2206.04901](#) (2022) 3
53. Liu, S., Zhang, X., Zhang, Z., Zhang, R., Zhu, J., Russell, B.: Editing conditional radiance fields. In: ICCV (2021) 3
54. Mascaro, R., Teixeira, L., Chli, M.: Diffuser: Multi-view 2D-to-3D label diffusion for semantic scene segmentation (2021) 24
55. Melas-Kyriazi, L., Laina, I., Rupprecht, C., Neverova, N., Vedaldi, A., Gafni, O., Kokkinos, F.: IM-3D: Iterative multiview diffusion and reconstruction for high-quality 3D generation. In: ICML (2024) 9
56. Melas-Kyriazi, L., Rupprecht, C., Laina, I., Vedaldi, A.: RealFusion: 360 reconstruction of any object from a single image. In: CVPR (2023) 1
57. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDEdit: Guided image synthesis and editing with stochastic differential equations. In: ICLR (2022) 3, 5, 9
58. Michel, O., Bar-On, R., Liu, R., Benaim, S., Hanocka, R.: Text2Mesh: text-driven neural stylization for meshes. In: CVPR (2022) 4
59. Mikaeili, A., Perel, O., Safaei, M., Cohen-Or, D., Mahdavi-Amiri, A.: SKED: sketch-guided text-based 3D editing. In: ICCV (2023) 3
60. Mildenhall, B., Srinivasan, P.P., Ortiz-Cayon, R., Kalantari, N.K., Ramamoorthi, R., Ng, R., Kar, A.: Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. ACM Transactions on Graphics (TOG) (2019) 9, 14, 26
61. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020) 1, 4
62. Mirzaei, A., Aumentado-Armstrong, T., Brubaker, M.A., Kelly, J., Levenshtein, A., Derpanis, K.G., Gilitschenski, I.: Watch your steps: Local image and scene editing by text instructions [abs/2308.08947](#) (2023) 3

63. Mirzaei, A., Aumentado-Armstrong, T., Derpanis, K.G., Kelly, J., Brubaker, M.A., Gilitschenski, I., Levinshtein, A.: Spin-nerf: Multiview segmentation and perceptual inpainting with neural radiance fields. In: CVPR. pp. 20669–20679 (2023) [3](#)
64. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: CVPR. pp. 6038–6047 (2023) [3](#)
65. Nguyen-Phuoc, T., Liu, F., Xiao, L.: SNeRF: stylized neural implicit representations for 3d scenes. SIGGRAPH **41**(4) (2022) [3](#)
66. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: GLIDE: towards photorealistic image generation and editing with text-guided diffusion models [abs/2112.10741](#) (2021) [3](#)
67. Park, J., Kwon, G., Ye, J.C.: Ed-nerf: Efficient text-guided editing of 3d scene using latent space nerf. In: ICLR (2024) [4](#)
68. Parmar, G., Singh, K.K., Zhang, R., Li, Y., Lu, J., Zhu, J.Y.: Zero-shot image-to-image translation. In: SIGGRAPH (2023) [3, 5](#)
69. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: improving latent diffusion models for high-resolution image synthesis [abs/2307.01952](#) (2023) [5](#)
70. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. In: ICLR (2023) [1, 2, 6, 10](#)
71. Qi, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., Chen, Q.: Fatezero: Fusing attentions for zero-shot text-based video editing. ICCV (2023) [2, 6](#)
72. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML. vol. 139, pp. 8748–8763 (2021) [3](#)
73. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022) [1, 3, 4, 5](#)
74. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: CVPR (2023) [3](#)
75. Sella, E., Fiebelman, G., Hedman, P., Averbuch-Elor, H.: Vox-e: Text-guided voxel editing of 3d objects. In: ICCV (2023) [2, 4, 6](#)
76. Shi, Y., Xue, C., Pan, J., Zhang, W., Tan, V.Y., Bai, S.: DragDiffusion: Harnessing diffusion models for interactive point-based image editing [2306.14435](#) (2023) [3](#)
77. Song, H., Choi, S., Do, H., Lee, C., Kim, T.: Blending-nerf: Text-driven localized editing in neural radiance fields. In: ICCV. pp. 14383–14393 (2023) [4](#)
78. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021) [1, 5](#)
79. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021) [8, 24](#)
80. Sun, C., Sun, M., Chen, H.: Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In: CVPR (2022) [2, 4](#)
81. Sun, C., Liu, Y., Han, J., Gould, S.: NeRFEeditor: differentiable style decomposition for full 3D scene editing. In: WACV (2022) [3](#)
82. Teng, Y., Xie, E., Wu, Y., Han, H., Li, Z., Liu, X.: Drag-a-video: Non-rigid video editing with point-based interaction [abs/2312.02936](#) (2023) [3](#)
83. Tretschk, E., Golyanik, V., Zollhöfer, M., Bozic, A., Lassner, C., Theobalt, C.: Scenerflow: Time-consistent reconstruction of general dynamic scenes [abs/2308.08258](#) (2023) [3](#)

84. Tschernezki, V., Laina, I., Larlus, D., Vedaldi, A.: Neural Feature Fusion Fields: 3D distillation of self-supervised 2D image representation (2022) [3](#)
85. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: CVPR (2023) [3, 8](#)
86. Vargas, F., Grathwohl, W.S., Doucet, A.: Denoising diffusion samplers. In: ICLR (2023) [5](#)
87. Wang, B., Dutt, N.S., Mitra, N.J.: ProteusNeRF: Fast lightweight NeRF editing using 3D-aware image context [abs/2310.09965](#) (2023) [4](#)
88. Wang, C., Chai, M., He, M., Chen, D., Liao, J.: CLIP-NeRF: Text-and-image driven manipulation of neural radiance fields. In: CVPR (2022) [4](#)
89. Wang, C., Jiang, R., Chai, M., He, M., Chen, D., Liao, J.: NeRF-Art: Text-driven neural radiance fields stylization [abs/2212.08070](#) (2022) [4](#)
90. Wang, D., Zhang, T., Abboud, A., Süsstrunk, S.: Inpaintnerf360: Text-guided 3d inpainting on unbounded neural radiance fields. arXiv preprint arXiv:2305.15094 (2023) [4](#)
91. Wang, J., Fang, J., Zhang, X., Xie, L., Tian, Q.: Gaussianeditor: Editing 3d gaussians delicately with text instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20902–20911 (2024) [4, 7](#)
92. Wang, X., Zhu, J., Ye, Q., Huo, Y., Ran, Y., Zhong, Z., Chen, J.: Seal-3D: Interactive pixel-level editing for neural radiance fields. In: ICCV (2023) [3](#)
93. Weder, S., Garcia-Hernando, G., Monszpart, A., Pollefeys, M., Brostow, G.J., Firman, M., Vicente, S.: Removing objects from neural radiance fields. In: CVPR. pp. 16528–16538 (2023) [3](#)
94. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7623–7633 (2023) [2, 6](#)
95. Xu, J., Wang, X., Cao, Y., Cheng, W., Shan, Y., Gao, S.: InstructP2P: Learning to edit 3D point clouds with text instructions [abs/2306.07154](#) (2023) [4, 6, 10](#)
96. Xu, S., Li, L., Shen, L., Lian, Z.: Desrf: Deformable stylized radiance field. In: CVPR. pp. 709–718 (2023) [3](#)
97. Xu, T., Harada, T.: Deforming radiance fields with cages. In: ECCV (2022) [3](#)
98. Yang, B., Bao, C., Zeng, J., Bao, H., Zhang, Y., Cui, Z., Zhang, G.: Neumesh: Learning disentangled neural mesh-based implicit field for geometry and texture editing. In: ECCV (2022) [3](#)
99. Yang, B., Zhang, Y., Xu, Y., Li, Y., Zhou, H., Bao, H., Zhang, G., Cui, Z.: Learning object-compositional neural radiance field for editable scene rendering. In: ICCV (2021) [3](#)
100. Yu, L., Xiang, W., Han, K.: Edit-diffnerf: Editing 3d neural radiance fields using 2d diffusion model. arXiv preprint arXiv:2306.09551 (2023) [4](#)
101. Yuan, Y., Sun, Y., Lai, Y., Ma, Y., Jia, R., Gao, L.: NeRF-editing: Geometry editing of neural radiance fields. In: CVPR (2022) [3](#)
102. Zhang, H., Feng, Y., Kulits, P., Wen, Y., Thies, J., Black, M.J.: Text-guided generation and editing of compositional 3d avatars. In: 2024 International Conference on 3D Vision (3DV) (2024) [4](#)
103. Zhang, J., Liu, X., Ye, X., Zhao, F., Zhang, Y., Wu, M., Zhang, Y., Xu, L., Yu, J.: Editable free-viewpoint video using a layered neural representation. ACM Trans. Graph. **40**(4) (2021) [3](#)
104. Zhang, K., Kolkin, N., Bi, S., Luan, F., Xu, Z., Shechtman, E., Snavely, N.: ARF: Artistic radiance fields. In: ECCV (2022) [3](#)

105. Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. In: NeurIPS (2023) 3
106. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023) 3
107. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018) 9
108. Zhang, S., Yang, X., Feng, Y., Qin, C., Chen, C.C., Yu, N., Chen, Z., Wang, H., Savarese, S., Ermon, S., Xiong, C., Xu, R.: Hive: Harnessing human feedback for instructional visual editing. In: CVPR (2024) 3
109. Zheng, C., Lin, W., Xu, F.: EditableNeRF: Editing topologically varying neural radiance fields by key points. In: CVPR (2023) 3
110. Zhou, S., Chang, H., Jiang, S., Fan, Z., Zhu, Z., Xu, D., Chari, P., You, S., Wang, Z., Kadambi, A.: Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21676–21685 (2024) 4
111. Zhou, X., He, Y., Yu, F.R., Li, J., Li, Y.: Repaint-nerf: Nerf editing via semantic masks and diffusion models. In: IJCAI (2023) 4
112. Zhuang, J., Wang, C., Lin, L., Liu, L., Li, G.: DreamEditor: Text-driven 3D scene editing with neural fields. In: SIGGRAPH (2023) 2, 4

Appendix

This supplementary material contains the following parts:

- **Additional Editing Results (Appendix A).** We provide additional qualitative examples of our method, including more editing prompts and partial editing results on complex scenes with many objects.
- **Additional Comparisons (Appendix B).** We provide in-depth comparisons of our approach and other update schemes and further comparisons with prior work, GaussianEditor.
- **Additional Details (Appendix C).** We provide more details regarding the experimental implementation and evaluation protocol.
- **Social Impact and Limitations (Appendix D).** We discuss the social impact and limitations of our work.

A Additional Editing Results

In this section, we provide additional qualitative editing results of DGE with more prompts and additional partial editing results on the Mip-NeRF360 [3] dataset.

Results with Additional Prompts. In Figure 7, we provide further qualitative editing results from our method, demonstrating the ability of DGE to generate complex and visually compelling editing effects. Specifically, when tasked with transforming the appearance of a human figure into mythical creatures,

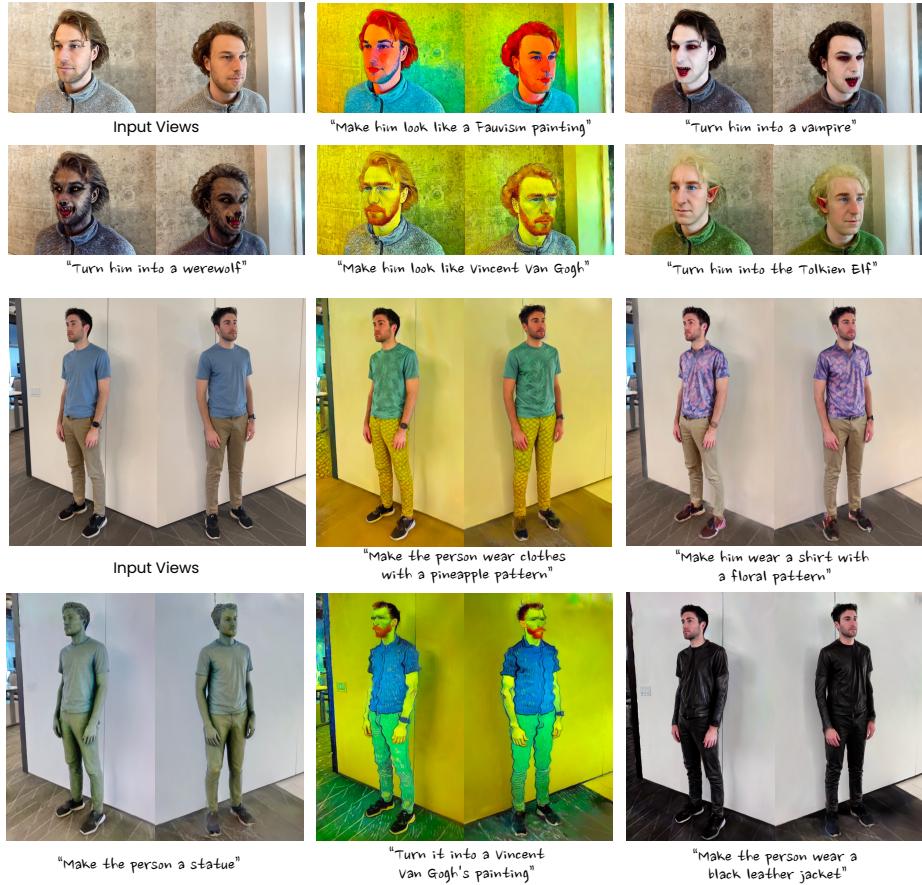


Fig. 7: Additional editing examples. We provide more qualitative results, including prompts to change the character or style. Our DGE produces realistic editing results with detailed textures.

such as a “*Vampire*,” “*Tolkien Elf*,” or “*Werewolf*,” DGE produces highly realistic results. This capability is not solely restricted to character modifications. We also show the ability of DGE to handle style transfer, such as converting the 3D models’ style into “*Fauvism painting*”. Finally, by directly and consistently editing multiple images at once, we show that DGE produces intricate 3D edits with detailed patterns and textures, *e.g.*, changing the clothes of the human figure to a “*pineapple pattern*” or “*floral pattern*”.

Results with Partial Editing. We provide additional partial editing results using three different scenes provided by Mip-NeRF360 [3] in Figure 8. Our DGE successfully edits the color of the selected objects such as *vase* and *dozer* while keeping other stuff unchanged. For example, only the vase in the first row is

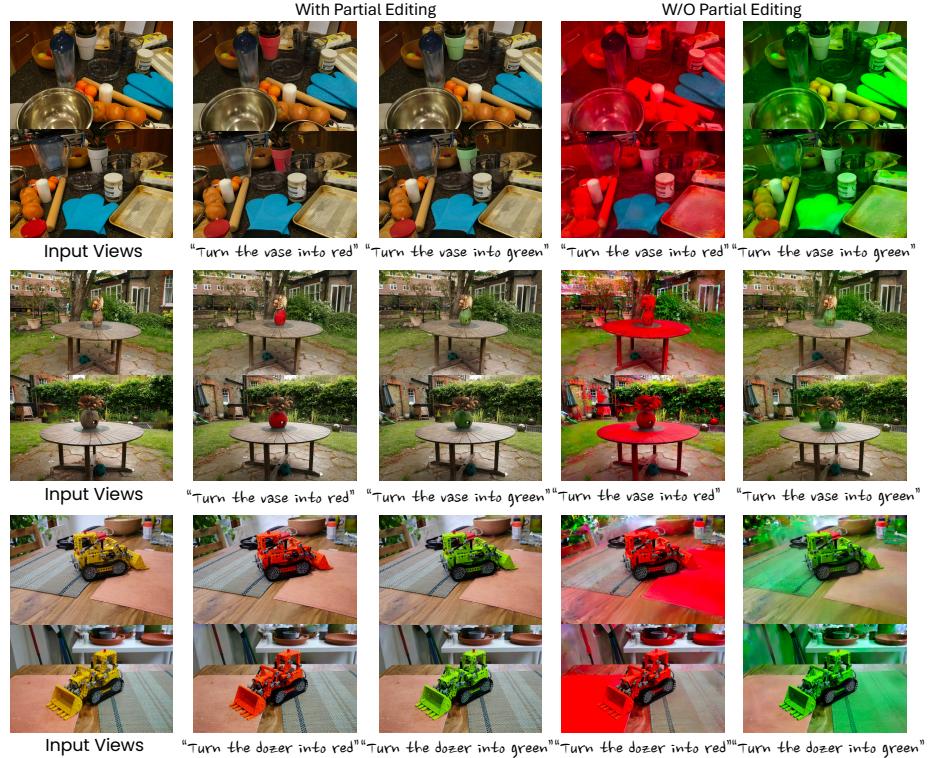


Fig. 8: Additional partial editing results. Our DGE successfully changes only the selected objects faithfully. We can even tell the color change through the transparent bowl before the vase.

changed to either green or red. Besides, we can still distinguish the color changes from the transparent bowl before the vase. As shown in the last two columns of the figure, without partial editing, the whole scene turns either red or green, which clearly demonstrates the advantages of partial editing. The core issue is the inability of the image editor (InstructPix2Pix) to isolate and modify the color of just the selected object, affecting the entire image instead.

B Additional Comparisons

Multi-view Consistent Editing vs. Iterative Dataset Updates. Next, we provide a qualitative comparison of our *direct* approach *with* and *without* multi-view consistency, as well as the iterative dataset update scheme. In Figure 9, we show the 2D edited images that are later used to update the corresponding 3D model. Note that all variants use the same base 3D model. We observe that when employing multi-view consistency in 2D editing, we achieve a uni-

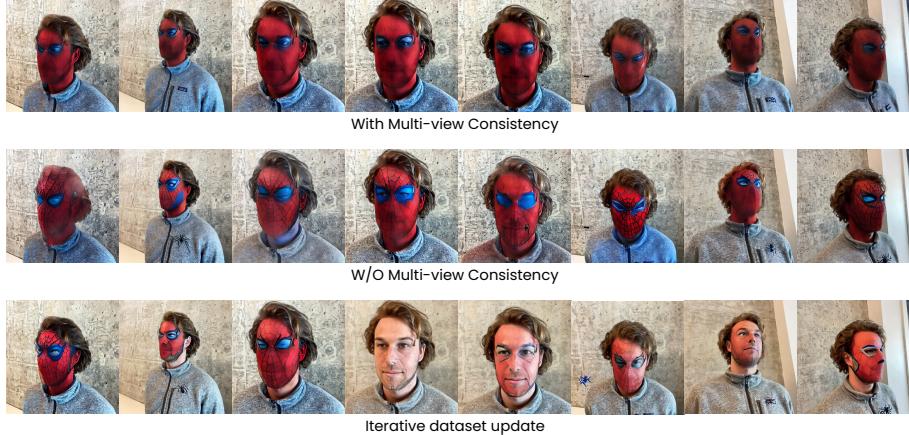


Fig. 9: Comparison of different updating methods. This figure compares the 2D editing results (i) with multi-view consistency, (ii) without multi-view consistency, and (iii) with iterative dataset updates. Multi-view consistent editing yields a consistent appearance across all different views.

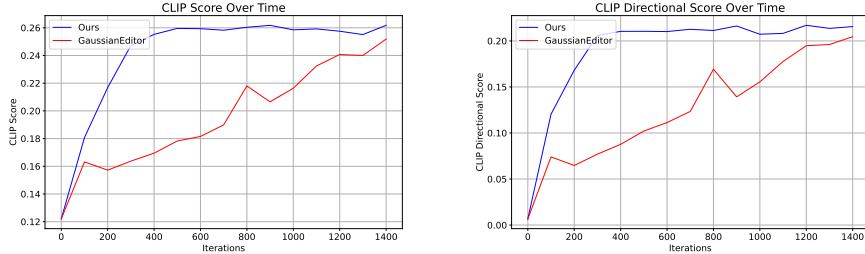


Fig. 10: CLIP and CLIP directional scores w.r.t. the number of iterations. We compare the CLIP scores of DGE and GaussianEditor over training iterations for the example depicted in Figure 9. Our method achieves high CLIP and CLIP directional scores in under 400 iterations, which is before the additional refinement step. We perform refining after 500 iterations.

form and consistent appearance across all images, which in turn accelerates the reconstruction process. Conversely, without multi-view consistency, the images are edited independently by InstructPix2Pixel. As a result, the appearance of the edited images tends to vary, despite the input views being similar (notably, the 3rd to 5th images in the middle row). Finally, the inconsistencies are exacerbated under the iterative dataset update scheme, which involves introducing different amounts of noise, resulting in considerable variability after editing. This variability contributes to the slower convergence speeds of methods such as IN2N [25] and GE [13].

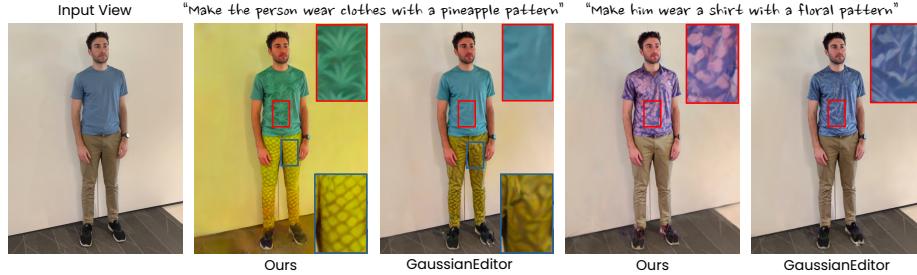


Fig. 11: Details Comparison. DGE generates editing results with more detailed textures such as the pineapple or floral pattern.

Extended Comparisons with GaussianEditor. GaussianEditor [13] is a recent work that attempts to edit Gaussian Splatting-based models similarly to IN2N [25]. Based on the iterative dataset update scheme, both GaussianEditor and IN2N iterate between 3D model updating and image editing. As discussed above, iterative dataset updates require many iterations to successfully apply the desired editing effect to the 3D model. In contrast, our DGE edits rendered images *simultaneously* and subsequently updates the Gaussian Splatting model, achieving faster speed and better quality. Figure 10 further demonstrates this by comparing the number of iterations required for convergence between DGE and GaussianEditor. By directly integrating consistent edits into the 3D model, we observe our approach that converges in significantly fewer iterations compared to GaussianEditor.

We provide an additional qualitative comparison with GaussianEditor to assess the edited image details in Figure 11. As shown in the figure, our method generates high-fidelity textures that align well with the textual description, which gives the human a pineapple-like outfit. In contrast, GaussianEditor falls short in delivering this complex pattern with adequate detail. This could be attributed to inconsistent editing results of InstructPix2Pix for different views during the process, as also previously noted by IN2N [25].

C Additional Details

C.1 Implementation Details

Image Editor. We use InstructPix2Pix as the base image editor. We follow the same setting in [13, 25], which uses DDIM [79] scheduler with 20 steps and the implementation included in the Diffusers [54] library. Different from IN2N [25] or GaussianEditor [13], which sample noise from $t \in [0.02, 0.98]$, we directly use $t = 1$ as we do not require iterative dataset updates. For the refinement step, we add noise with $t = 0.3$ to capture more details of 3D editing. We select one keyframe in every five frames.



Fig. 12: The effect of camera sorting. The first two rows for each scene are the ground-truth images of sorted cameras and the bottom two rows are ground-truth images of unsorted cameras. Note the smoother transitions from one frame to the next after sorting.

Camera Sorting. Our approach to multi-view is inspired by video editing. While in videos camera trajectories are typically smooth, in our case, when rendering views from the 3D model, we do not assume that a camera trajectory is known in advance. However, obtaining such a trajectory is useful to ensure that the selected keyframes will cover a variety of viewpoints. To this end, prior to multi-view editing, we sort the rendered viewpoints as follows. (1) We take the camera with the largest value along the x-axis in world coordinates as the reference camera. (2) We sort the cameras using the relative angle between their forward vectors and that of the reference camera. The forward vector is defined as the normalized vector along the z-axis of the camera in world coordinates.

In Figure 12, we visualize the images corresponding to sorted cameras and unsorted cameras for two different scenes.

Computation Time. The average editing time reported (for both DGE and other methods) is evaluated on one Nvidia RTX A6000 GPU with 48G memory.

Scene	Source Prompt	Target Prompt	Edit Instruction
Face	“A man with curly hair in a grey jacket”	“A man with curly hair in a grey jacket with a Venetian mask”	“Give him a Venetian mask”
Face	“A man with curly hair in a grey jacket”	“A man with curly hair in a checkered cloth”	“Give him a checkered jacket”
Face	“A man with curly hair in a grey jacket”	“A spider man with a mask and curly hair”	“Turn him into spider man with a mask”
Bear	“A stone bear in a garden”	“A robotic bear in the garden”	“Make the bear look like a robot”
Bear	“A stone bear in a garden”	“A panda in the garden”	“Make the bear look like a panda”
Bear	“A stone bear in a garden”	“A bear with rainbow color”	“Make the color of the bear look like rainbow”
Person	“A man standing next to a wall wearing a blue T-shirt and brown pants”	“A man looks like a mosaic sculpture standing next to a wall”	“Make the man look like a mosaic sculpture”
Person	“A man standing next to a wall wearing a blue T-shirt and brown pants”	“A man wearing a shirt with a pineapple pattern”	“Make the person wear a shirt with a pineapple pattern”
Person	“A man standing next to a wall wearing a blue T-shirt and brown pants”	“An Iron Man stands to a wall”	“Turn him into Iron Man”
Person	“A man standing next to a wall wearing a blue T-shirt and brown pants”	“A robot stands to a wall”	“Turn the man into a robot”

Table 3: Detailed evaluation dataset. As shown above, we evaluate different methods on the above 10 different scene-prompt pairs. The source and target prompts are used for computing CLIP and CLIP directional scores. The edit instructions are textual inputs of InstructPix2Pix [4].

C.2 Evaluation Prompts

We have created a set of 10 scene-prompt pairs for evaluation purposes, shown in Table 3. These are used for computing the CLIP and CLIP directional scores. We randomly sample 20 views from the training cameras and use the average score as the score for each scene-prompt pair.

D Social Impact and Limitations

Social Impact. In this work, we leverage existing datasets in previous studies [3, 25, 60]. Our examples may include human subjects for whom consent has

been obtained, as outlined in [25]. The potential outcomes of our editing results may lead to negative societal impacts if utilized for harmful purposes. We urge users to employ our methodology responsibly, adhering to ethical guidelines, regulations, and relevant laws.

Limitations. The editing performance of our approach is primarily constrained by the capabilities of the underlying image editor, in this case, InstructPix2Pix [4]. Consequently, it inherits limitations from InstructPix2Pix. As such, our method’s capability to handle substantial geometric transformations is limited, such as adjusting the position of a person’s arm or simulating actions like jumping. We anticipate that incorporating advanced or alternative 2D diffusion priors could pave the way for overcoming these constraints in the future.