

Control4D: Efficient 4D Portrait Editing with Text

Ruizhi Shao¹, Jingxiang Sun¹, Cheng Peng¹, Zerong Zheng^{1,2}, Boyao Zhou¹, Hongwen Zhang¹, Yebin Liu¹

¹Department of Automation, Tsinghua University ²NNKosmos Technology



Figure 1. We propose Control4D, an approach to high-fidelity and spatiotemporal-consistent 4D portrait editing with only text instructions. Given the multi-view videos as shown in the left and text instructions "Jensen Huang is roasting steak", Control4D generates realistic and 4D consistent editing results presented in the middle and right.

Abstract

We introduce Control4D, an innovative framework for editing dynamic 4D portraits using text instructions. Our method addresses the prevalent challenges in 4D editing, notably the inefficiencies of existing 4D representations and the inconsistent editing effect caused by diffusion-based editors. We first propose GaussianPlanes, a novel 4D representation that makes Gaussian Splatting more structured by applying plane-based decomposition in 3D space and time. This enhances both efficiency and robustness in 4D editing. Furthermore, we propose to leverage a 4D generator to learn a more continuous generation space from inconsistent edited images produced by the diffusion-based editor, which effectively improves the consistency and quality of 4D editing. Comprehensive evaluation demonstrates the superiority of Control4D, including significantly reduced training time, high-quality rendering, and spatial-temporal consistency in 4D portrait editing. The link to our project website is: <https://control4darxiv.github.io/>.

1. Introduction

The realm of 4D scene reconstruction has witnessed advancements with the advent of dynamic neural 3D representation [12, 35, 47, 51, 59]. These innovations have significantly enhanced our ability to capture and represent dynamic scenes. Despite these advances, the interactive editing of these 4D scenes still poses substantial

challenges. The primary challenge involves ensuring both spatial-temporal consistency and high quality in 4D editing.

Available 4D editing techniques [26, 52], while effective for fundamental tasks like object removal or color modification, often fall short in delivering interactive and advanced editing functionalities. Recently, a groundbreaking framework based on text-to-image (T2I) diffusion model [50] has emerged for 3D generation and editing. It integrates a neural 3D representation such as NeRF with an image diffusion model and achieves text-to-3D generation [8, 33, 39, 50] or editing [17] by iteratively aligning images rendered from the 3D representation with those generated by the diffusion model. This diffusion-based framework allows for more flexible and enhanced editing through textual control.

Building on this framework, a straightforward approach to 4D editing involves transitioning from a 3D to a 4D representation. However, it faces two primary challenges: First, 4D representations such as dynamic NeRFs require dense sampling along the rays to render images, which is slow and highly memory-intensive [12, 51, 59]. Such inefficiency significantly increases the time required for editing in 4D scenarios. On the other hand, current T2I diffusion models lack consistency in editing different images [17]. This inconsistency is more apparent in 4D editing, as the results vary across different spatial perspectives and over time, making 4D editing extremely challenging.

In this paper, we address these challenges and present Control4D, a novel method for efficient, high-quality, and consistent 4D dynamic portrait editing with text as input. Firstly, to enhance the efficiency of 4D representation, we

propose to extend an explicit 3D representation, Gaussian Splatting, to a 4D dynamic representation. Gaussian Splatting is an emerging representation that has demonstrated its efficiency in training and rendering for 3D reconstruction [27] and generation [67]. However, as it uses discrete Gaussian point clouds where every point is independent from each other, it easily introduces noise during the 4D editing process, where the edited images are not consistent in both space and time. To address this issue, we first propose to construct the spatial structure to describe the attributes of discrete Gaussian points by a unified, structured tri-plane [7] representation. Specifically, we project each Gaussian point onto three feature planes and employ an MLP to integrate features and derive their attributes, which not only ensures efficiency but also enhances robustness. Then, we extend Gaussian Splatting to 4D by defining a canonical Gaussian point cloud and allowing each point to move with time. To regularize the flow of discrete points, we also project their positions with time into 9 planes [59] to make the flow more structured. With the tri-planar structure for the canonical space and the 4D plane-based structure for the 4D flow, we introduce GaussianPlanes representation, which significantly reduces the time cost and improves spatiotemporal consistency in 4D editing.

Although GaussianPlanes significantly improves the efficiency of representation, implementing 4D editing based on it still encounters a bottleneck. This bottleneck lies in the T2I diffusion model, as the diffusion-based editor adopts a 2D generation process and produces inconsistent edits in 4D space across time and viewpoints. Consequently, when optimized with these inconsistent images, the dynamic scene model tends to diverge or produce blurry and smoothed outcomes. To overcome this challenge, we propose a 4D generator to mitigate the issue of inconsistent supervision arising from the edited dataset. The key insight of our method is to learn a more continuous GAN latent space based on the edited images produced by the editor, avoiding direct but inconsistent supervision. Specifically, we introduce additional latent properties to GaussianPlanes and incorporate it with a 2D super-resolution module, constructing a 4D generator, capable of producing high-resolution images based on the rendered latent features. Simultaneously, we employ a discriminator to learn the generation distribution from the edited images, which then provides discrimination signals for updating the generator. To ensure stable training, we extract multi-level information from the edited images and utilize it to facilitate the generator’s learning process.

We conduct comprehensive evaluation of our approach using a diverse collection of dynamic portraits. To validate the efficacy of our design, we conduct ablation studies and compare our method with a 4D extension of Instruct-NeRF2NeRF [17]. The evaluation demonstrates the efficiency and remarkable capabilities of our method in achiev-

ing both photo-realistic rendering and spatio-temporal consistency in 4D portrait editing. To sum up, our main contributions are listed as follows:

- We propose an efficient and robust 4D representation GaussianPlanes for 4D editing by applying plane-based decomposition to structure Gaussian Splatting in both space and time.
- We introduce a 4D generator to learn from the 2D diffusion-based editor, which reduces the effect of inconsistent supervision signals and enhances the quality of 4D editing.
- Building upon the proposed GaussianPlanes and 4D generator, We introduce Control4D, a novel framework for flexible 4D portrait editing with text, which significantly reduces the training time, achieves high-quality rendering, and ensures spatio-temporal consistency.

2. Related Work

2.1. 2D Diffusion Models

Diffusion models iteratively transform random samples into ones resembling target data [9, 20, 29, 64]. Enhanced with pre-trained models [53], they solve multi-modal tasks like text-to-image generation [19, 44, 54]. VQdiffusion[15] and LDMs [55] bolster performance by operating within an autoencoder’s latent space. Although these models have found success, temporally inconsistent issues emerge in videos and 4D scenes.

Research has also concentrated on diffusion-based video generation and editing. Video Diffusion Models (VDM)[22] use U-Net architecture to train image and video data jointly, while approaches like ImagenVideo[21] enable high-resolution video generation. Various methods aim to transfer text-image generation to text-video, but due to training costs, many focus on text-prompted video editing [2, 6, 11, 28, 31, 62, 79, 86]. These efforts underscore the potential of text-based video editing, yet challenges related to temporal consistency, quality generation, and viewpoint alterations persist.

2.2. NeRF-Based 3D Generation and Editing

NeRFs [40] have gained widespread popularity for producing realistic 3D scene reconstruction and novel views based on calibrated photographs, and have been further developed in numerous subsequent studies [69]. Nevertheless, NeRFs still pose a challenge for editing purposes, primarily due to their underlying representation.

NeRF editing researchers have focused on utilizing GANs [13] and Diffusion models for their powerful generative capabilities. GAN-based methods have seen a proliferation of novel architectures that combine implicit or explicit 3D representations with neural rendering techniques, achieving promising results [7, 14, 42, 43, 78, 87]. How-

ever, voxel-based GANs face challenges such as high memory requirements and computational burden when training high-resolution 3D GANs. On the other hand, diffusion-based methods have two primary approaches for extending 2D editing to 3D NeRFs. The first involves using Stable Diffusion with score distillation sampling (SDS) loss to generate 3D NeRFs using the 2D diffusion-prior, as seen in DreamFusion [50] and its follow-ups [5, 8, 24, 32, 33, 37, 49, 57, 60, 61, 63, 66–68, 72, 75]. However, these methods can only generate isolated objects lacking fine-level control over synthesized outputs. The second approach utilizes dataset update(DU) to guide NeRF convergence iteratively, as seen in Instruct-NeRF2NeRF [17], but it has network convergence issues and can be cost-intensive.

2.3. NeRF for Dynamic Scenes

To expand the success of NeRF into the temporal domain, researchers have pursued the strategy of modeling scenes in 4D domain with time dimension. DyNeRF [30] proposes a keyframe-based training strategy to extend NeRF with time-conditioning. VideoNeRF [80] learns a spatiotemporal irradiance field directly from a single video and resolves the shape-motion ambiguities in monocular inputs by incorporating depth estimation. Meanwhile, NeRFlow [10] and DCT-NeRF[71] utilize point trajectories to regularize network optimization. Park et al.[46, 47]; Pumarola et al. [51]; Tretschk et al. [70] adopt a similar framework that introduce a separate MLP to predict scene deformations for multi-view and monocular videos, respectively. Another approach for dynamic scenes is DeVRF [35], which adopts a voxel-based representation to model both the 3D canonical space and the 4D deformation field. Additionally, methods including Neuralbody [48] and [36, 76, 84, 85] leverage parametric body templates as semantic priors to achieve photo-realistic novel view synthesis of complex human performances. Recently, to achieve higher quality with lower memory, NeRFPlayer [65] have decomposed the 4D space into regions of static, deforming, and newly appeared content. Meanwhile, more compact and efficient representations, such as [4, 12, 25, 81] are proposed, significantly boosting the rendering quality and efficiency.

2.4. Gaussian Splatting

3D Gaussian Splatting (3DGS) [27] offers a high-quality, swift alternative to Neural Radiance Fields (NeRF), leveraging differentiable 3D Gaussians for efficient rasterization. Unlike NeRF and other implicit 3D representations [45, 73] which render images based on volume rendering, 3D-GS employs a splatting method for image rendering, resulting in real-time speed. The successor, 4D Gaussian Splatting (4DGS) [38, 77], extends this with per-frame dense tracking and novel view synthesis for dynamic scenes by utilizing a lightweight deformation field to model Gaussian motions

and shape changes.

3. Overview

To achieve high-quality, efficient, and consistent 4D portrait editing, we first extend the Gaussian Splatting to 4D representation and structure it through a spatial-temporal plane-based decomposition (Sec. 4). To address the issue of inconsistencies in edited images generated by diffusion-based editors, we integrate a 4D Editor with GaussianPlanes to effectively mitigate instability and blurring issues and achieve the realism and quality of 4D editing (Sec. 5). Building on the GaussianPlanes and 4D Editor, we finally introduce several efficient training strategies for 4D editing in the Control4D framework (Sec. 5.3).

As shown in Fig. 2, our framework consists of the GaussianPlanes and a 4D generator. Given multi-view videos, we first reconstruct the 4D portrait based on GaussianPlanes. Subsequently, we edit the reconstructed rendering results and latent features through a multi-level generator to obtain the edited results. Simultaneously, we employ an iterative approach to achieve dataset update through a 2D diffusion-based editor, which is a ControlNet [83] in practice. The outputs of the editor serve as real images, while the generator’s results function as fake images for the discriminator’s input. As the GAN training progresses, we progressively incorporate the generator’s outputs to refine the inputs of the 2D diffusion-based editor, facilitating training convergence. Ultimately, the discrimination outcomes are utilized to compute the GAN loss, driving the iterative refinement of both the generator and discriminator. This methodology ensures the efficient and precise realization of 4D editing through our GAN-based framework.

4. GaussianPlanes

In this section, we propose an efficient and robust 4D representation GaussianPlanes for 4D portrait editing. The key idea is to structure the discrete points of Gaussian Splatting in both space and time. First, we introduce the spatial tri-plane decomposition, which makes Gaussian Splatting structured in the spatial domain (Sec. 4.1). Following this, we expand Gaussian Splatting into 4D representation and structure the flow of each Gaussian point by performing a temporal-spatial plane-based decomposition (Sec. 4.2).

4.1. GaussianPlanes in 3D

Gaussian Splatting is an emerging explicit 3D representation that utilizes a set of Gaussian point clouds to represent 3D scenes. Each point is described with attributes of the center position $\mathbf{x} \in \mathbb{R}^3$, the rotation quaternion $\mathbf{r} \in \mathbb{R}^4$, the scale factor $\mathbf{s} \in \mathbb{R}^3$, the opacity value $\alpha \in \mathbb{R}$ and the color feature $\mathbf{c} \in \mathbb{R}^3$. The rendering process of Gaussian Splatting involves projecting the Gaussian point cloud

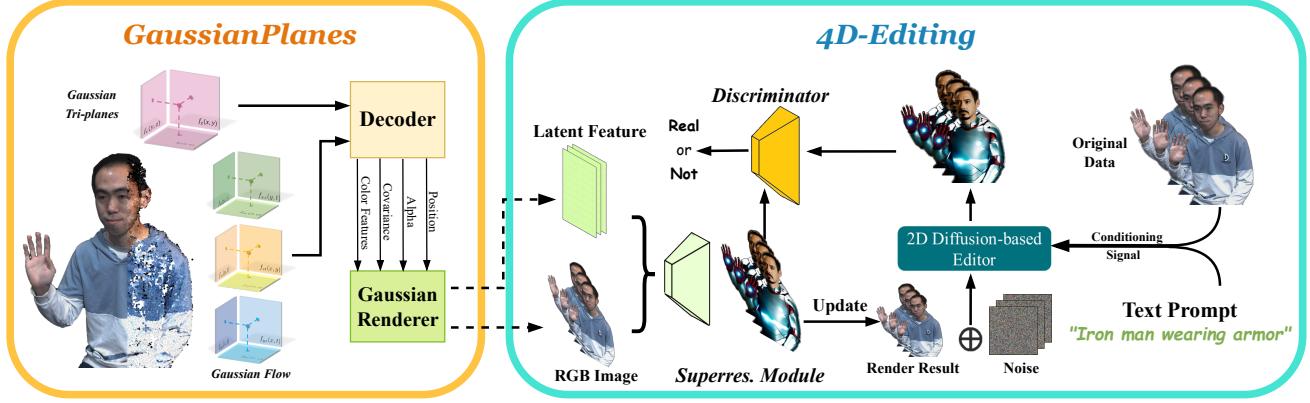


Figure 2. **Pipeline of Control4D:** Our method first utilizes GaussianPlanes to train the implicit representation of a 4D portrait scene, which are then rendered into latent features and RGB images using Gaussian rendering, serving as inputs for the GAN-based generator. Meanwhile, we apply the 2D-diffusion-based editor to edit the dataset with the noisy results and conditions as inputs, leading to updated results that are used as real images while the Superres. Module’s outputs serve as fake images fed into the Discriminator for discrimination. The discriminative results are used to calculate loss, allowing for iterative updates of both the Generator and Discriminator.

onto the rendering viewpoint according to camera parameters, followed by rasterization and volume rendering. Since each point in the Gaussian point cloud is independent and unstructured, noise easily occurs during optimization. To enhance robustness, we propose a spatial tri-plane decomposition to represent the attributes of the Gaussian points. Specifically, we decompose the color \mathbf{c}_i , opacity α_i and rotation \mathbf{r}_i of i -th Gaussian point into tri-plane features:

$$\begin{aligned} \mathbf{c}_i &= f_{\mathbf{c}}(F_{\mathbf{c}}^{xy}(x_i, y_i), F_{\mathbf{c}}^{xz}(x_i, z_i), F_{\mathbf{c}}^{yz}(y_i, z_i)), \\ \alpha_i &= f_{\alpha}(F_{\alpha}^{xy}(x_i, y_i), F_{\alpha}^{xz}(x_i, z_i), F_{\alpha}^{yz}(y_i, z_i)), \\ \mathbf{r}_i &= f_{\mathbf{r}}(F_{\mathbf{r}}^{xy}(x_i, y_i), F_{\mathbf{r}}^{xz}(x_i, z_i), F_{\mathbf{r}}^{yz}(y_i, z_i)), \end{aligned} \quad (1)$$

where F^{xy} , F^{xz} , F^{yz} are the decomposed feature planes, and f is an MLP that fuses features to predict specific attributes. In this way, although the Gaussian points remain independent, their attributes are structured and low-rank in spatial space, which helps to reduce noise and improve the robustness of Gaussian Splatting. The scale factor s and center position \mathbf{x} are not decomposed, as splitting Gaussian points would abruptly halve the scale factor and the center position of each point is used for querying attributes itself.

4.2. GaussianPlanes in 4D

To extend Gaussian Splatting for 4D editing, we regard the Gaussian point cloud at the first frame as the canonical space and represent the 4D scene at different times by deforming the canonical Gaussian point cloud. Specifically, we define the flow $\hat{\mathbf{x}}, \hat{\mathbf{r}}$ for both position and rotation attributes of Gaussian points. Then, for time t , we move each Gaussian point in the canonical space ($t = 0$) with the flow:

$$\begin{aligned} \mathbf{x}_i(t) &= \mathbf{x}_i(0) + \hat{\mathbf{x}}_i(t), \\ \mathbf{r}_i(t) &= \mathbf{r}_i(0) + \hat{\mathbf{r}}_i(t). \end{aligned} \quad (2)$$

In this way, we enhance temporal consistency since the Gaussian point cloud at all times corresponds to its canonical space. However, the flow of each gaussian point is still discrete and independent. To further structure the flow of Gaussian points in space and time, we adopt spatial-temporal plane-based decomposition proposed by Tensor4D [59] and decompose the flow attributes of i -th point into nine feature planes:

$$\begin{aligned} \hat{\mathbf{x}}_i(t) &= f_{\hat{\mathbf{x}}}(x_i, y_i, z_i, t) = \pi_3(\Pi_3(F_{\hat{\mathbf{x}}})), \\ \hat{\mathbf{r}}_i(t) &= f_{\hat{\mathbf{r}}}(x_i, y_i, z_i, t) = \pi_3(\Pi_3(F_{\hat{\mathbf{r}}})) \end{aligned} \quad (3)$$

where Π_3, π_3 are the hierarchical 4D decomposition in Tensor4D and F represents feature planes. Through spatial tri-planar decomposition and 4D plane-based decomposition, we structure the 4D Gaussian Splatting to enhance its consistency while maintaining efficiency.

5. 4D Editing with GaussianPlanes

To solve another challenge raised by diffusion-based editors, we propose a GaussianPlane-based 4D generator to edit 4D scene from the 2D inconsistent editing images with stable optimization. Instead of utilizing direct supervision with the edited images [3], our method learns a continuous generation space via GAN [13] to establish a connection between GaussianPlanes and dynamically edited images. Specifically, we integrate GaussianPlanes with a 2D GAN-based super-resolution module into a 4D generator and learn a generation space from the edited images generated by the diffusion model. Leveraging its generative capabilities, the 4D generator can effectively distill knowledge from the diffusion-based editor and distinguish between the rendering images (fake samples) and edited images (real samples). Subsequently, GaussianPlanes can be

optimized within a continuous generative space supervised by the discrimination loss. With such a learning-to-generate mechanism, our method effectively alleviates blurry effects, resulting in high-fidelity and consistent 4D editing. In the following, we will introduce 1) integrating GaussianPlanes with GAN for 4D scene generation; and 2) the generation with multi-level guidance.

5.1. Connecting GAN to GaussianPlanes

To enable the generative ability of GaussianPlanes and higher rendering resolution, we build a 4D generator by connecting the GaussianPlanes representation with a GAN-based super-resolution module. To this end, we first augment each point in the GaussianPlanes with latent features [7] as additional attributes. Here we assume the latent features follow a normal distribution, so in practice we augment the Gaussian attributes with their means μ and variances σ , thus enabling subsequent sampling. We also adopt the same tri-plane decomposition for these latent distribution parameters. Then we can render a “distribution parameter map” to sample the latent features, which will be fed into a super-resolution module G . Meanwhile, we also render an RGB image for auxiliary supervision. The distribution map consist of a latent mean map and a latent variance map, denoted as \mathbf{I}_μ and \mathbf{I}_σ , respectively, which capture the mean and variance of latent features. By leveraging this distribution map, we then proceed to sample a latent feature map that will be fed into G :

$$\mathbf{I}_l = \mathbf{I}_\mu + t\mathbf{I}_\sigma, t \sim N(0, 1). \quad (4)$$

Then, we concatenate the rendered RGB images I_r and the latent feature maps I_l and feed them into the super-resolution module to synthesize high-resolution images:

$$\mathbf{I}_G = G(\mathbf{I}_r, \mathbf{I}_l). \quad (5)$$

As mentioned above, the edited images are temporally inconsistent due to the frame-by-frame editing. To avoid the discrete and inconsistent issue of direct supervision, our method learns a more continuous generation space via GAN from these edited images. Specifically, the generated images \mathbf{I}_G are considered as fake samples, while the edited images are regarded as real samples. The GAN loss can be formulated as follows:

$$\begin{aligned} L_D &= D(\mathbf{I}_G) - D(\mathbf{I}_{ed}) + L_{gp} \\ L_G &= -D(\mathbf{I}_G), \end{aligned} \quad (6)$$

where \mathbf{I}_{ed} are the edited images generated by diffusion-based editor, D is the discriminator and L_{gp} is the Wasserstein GAN gradient penalty loss [1].

5.2. Multi-level Generation with Guidance

When training GAN with the loss in Eqn. 6, we observe that the learning process often suffers from mode collapse

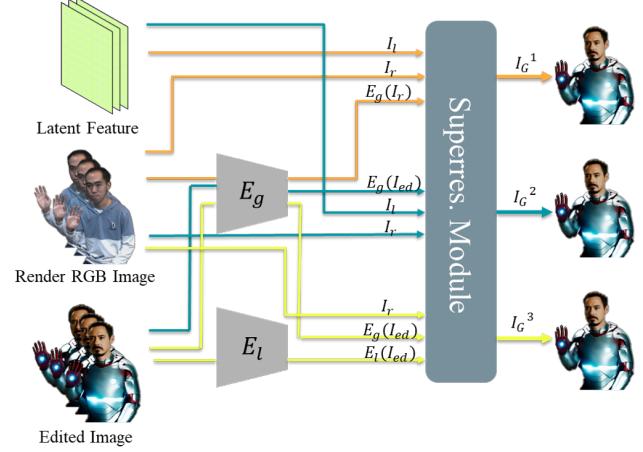


Figure 3. **Illustration of the Generation with Multi-level Guidance:** we propose a three-level image generation process to balance the generator training, where E_g denotes for the global encoder and E_l denotes for the local encoder.

issue. This may be caused by the fact that there is a limited number of edited images, and it is easy for the discriminator to learn how to distinguish between different sources of samples. To stabilize the learning process, we propose to extract multi-level information from the edited images and use these global and local cues to guide the learning of the generator. As shown in Fig. 3, during the training process, we construct two networks—global encoder E_g and local encoder E_l —to extract the global code and local feature maps of the edited image \mathbf{I}_{ed} , respectively. With these conditions as additional inputs, our generator can synthesize images on three levels:

$$\begin{aligned} \mathbf{I}_G^1 &= G(\mathbf{I}_r, \mathbf{I}_l, E_g(\mathbf{I}_r)) \\ \mathbf{I}_G^2 &= G(\mathbf{I}_r, \mathbf{I}_l, E_g(\mathbf{I}_{ed})) \\ \mathbf{I}_G^3 &= G(\mathbf{I}_r, E_l(\mathbf{I}_{ed}), E_g(\mathbf{I}_{ed})) \end{aligned} \quad (7)$$

Throughout the progression from level 1 to level 3, the generator produces images that gradually approach real edited images:

- At level 1, the generator directly synthesis images based on Tensor4D.
- At level 2, global information from the real edited images is introduced as conditions, guiding the generator to produce results consistent with the overall style of the real images.
- At level 3, both the global and local information from the real edited images is used as conditions, enabling the network to generate images that exhibit consistency in both the overall pattern and finer details with the real edited images.

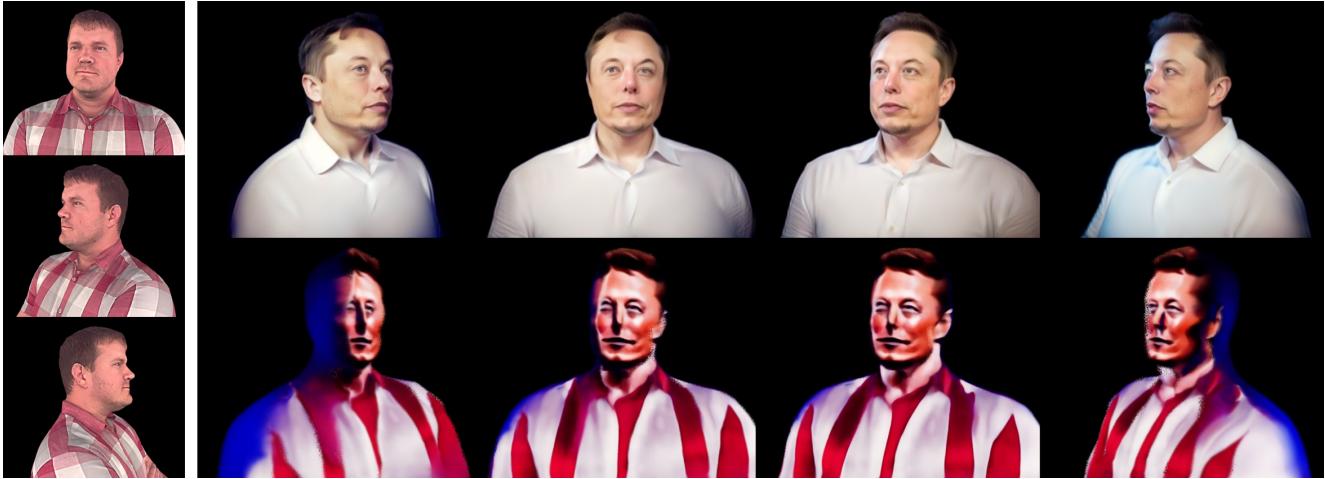


Figure 4. ***Qualitative comparisons with Instruct-NeRF2NeRF(static)***: In a static scenario, given the prompt “Turn him into Elon Musk”, train the model to converge and we can see that, on the same dataset, our method (the top row) produces highly realistic renderings of human portraits, while instruct nerf2nerf exhibits lower levels of realism and consistency, along with unexpected distortions in facial features.

To facilitate training, we also utilize different losses at different levels:

$$\begin{aligned} L_1 &= -D(\mathbf{I}_G^1) \\ L_2 &= -D(\mathbf{I}_G^2) + L_P(\mathbf{I}_G^2, \mathbf{I}_{ed}) \\ L_3 &= -D(\mathbf{I}_G^3) + L_P(\mathbf{I}_G^3, \mathbf{I}_{ed}) + \|\mathbf{I}_G^3 - \mathbf{I}_{ed}\|_1 \end{aligned} \quad (8)$$

Level 1 employs the original GAN loss. At level 2, a perceptual loss is introduced as an additional constraint to enforce consistency in the global style. At level 3, the loss function simultaneously incorporates L1 loss, perceptual loss, and GAN loss as penalties, as the consistency in details and global style is desired. This multi-level information guides the generator to converge progressively towards the generation space of the diffusion model, improving training stability in single scenarios and accelerating convergence compared to the original GAN training process.

5.3. Training Strategy

To address the high iterative optimization cost associated with using the diffusion-based editor, we propose several strategies to further improve the efficiency of 4D editing.

Staged Training Strategy. We adopt a staged training strategy that facilitates convergence. First, we fix the flow in the static stage and focus solely on editing the canonical space. This simplifies the editing process from 4D to 3D static editing, resulting in faster convergence. Once the editing of the canonical space has converged, we proceed to train GaussianPlanes across the entire 4D sequences. We also adopt a smaller noise timestep $t \in U(0.02, 0.6)$ for the diffusion-based editor in the dynamic stage since most of the editing effect is done in the static stage.

Batch-based Dataset Update. To improve the editing consistency across different images, instead of editing a sin-

gle image per iteration like InstructN2N, we group several images as a batch and edit them simultaneously. In editing each batch, we incorporate an attention module [16] for multi-frame image generation into our diffusion-based editor to capture the temporal-spatial correspondences and thereby improve editing consistency.

6. Experiment

We primarily conduct experiments on the dynamic Tensor4D dataset, which captures dynamic half-body human videos by four sparsely positioned, fixed RGB cameras. The calibration is performed using a checkerboard. Each data sample captures a diverse range of human motions in 1-2 minute duration. For our experiments, we extract 2-second segments, consisting of 50 frames, from the full-length videos for 4D reconstruction and editing. Furthermore, to showcase the capabilities of our method in 360-degree scenes, we also select scanned human models from Twindom [82] dataset for additional evaluation. Please refer to the suppl. for more experiment details.

6.1. Qualitative Evaluation

6.1.1 Static scene

Since the task of 4D editing with text has not been addressed in previous works, we first conduct evaluation on static scene in order to validate the efficiency of our proposed methods. To validate the efficiency of our proposed GAN, we first conduct a comparison between NeRF+GAN and instruct-NeRF2NeRF under static scenes. We select some human models from the Twindom dataset and sampled 180 viewpoints randomly within a 360-degree range to render images. Subsequently, we evaluate NeRF+GAN and



Figure 5. *Qualitative comparisons with baseline(dynamic)*: In a dynamic scenario, given the prompt “Mark Zuckerberg”, compared to the baseline result (the first row) that only employs the dataset update (DU) method, our proposed approach (with the addition of GAN, the second row) demonstrates higher levels of realism and consistency in our rendered results.

Method	FID \downarrow	CLIP Similarity \uparrow
Static scene		
InstructPix2Pix	-	0.3089
ControlNet	-	0.3313
InstructNeRF2NeRF	126.3	0.2989
Tensor4D	118.7	0.3316
Tensor4D+GAN	27.81	0.3334
GaussianPlanes	49.32	0.3301
Control4D (Ours)	14.11	0.3323
Dynamic scene		
ControlNet	-	0.3185
Tensor4D	155.6	0.3144
Tensor4D+GAN	47.39	0.3178
GaussianPlanes	67.58	0.3175
Control4D (Ours)	18.59	0.3192

Table 1. Quantitative Comparisons on static and dynamic scenes.

instruct-NeRF2NeRF for editing with prompt “Turn him into Elon Musk”. In Fig. 4, we present the results after 50,000 iterations of training. Observing the results, it is evident that our GAN can generate images of high quality, exhibiting rich detail and enhanced realism. In contrast, the instruct-NeRF2NeRF outputs appear smoother, with some issues observed in the blending of side views. This comparison highlights the significant advantage of our GAN in terms of editing capabilities.

6.1.2 Dynamic scene

In dynamic scenarios, we compare our proposed method and the baseline method that only utilize GaussianPlanes, and the results are presented in Figure 5. We also present the results of different individuals engaged in various actions, which can be referenced in Figure 6. In the baseline approach, where GAN-based generation is not utilized, GaussianPlanes is directly tasked with fitting a dynamically changing editing dataset in both space and time. This direct fitting process often leads to the optimization of smooth results that may lack consistency and high-quality details. Our proposed method incorporates GAN-based generation, leveraging the GAN to learn a more continuous 4D generation space. This allows us to leverage the smooth supervisory signals for optimization. Thus, our method generates consistent and high-quality results that exhibit improved fidelity and capture finer details in the dynamic editing process. The comparison between the baseline approach and our method demonstrates the effectiveness of our proposed 4D generator in enhancing the overall quality and consistency of the generated results.

6.2. Quantitative Experiment

We conducted quantitative experiments in 5 static and 4 dynamic scenarios. The results are presented in Tab. 1. First, we compare the diffusion-based editor including Instructpix2pix [3] and ControlNet [83] in the context of portrait editing. ControlNet exhibited better consistency between the subject and the editing prompt than Instruct-



Figure 6. Qualitative results on Tensor4D dataset. Our method produces coherent and realistic outcomes that maintain high fidelity and accurately preserve intricate details throughout the dynamic editing procedure. The prompts we use are "Emma Watson", "Taylor Swift", "Iron Man wearing armor", "Captain Jack Sparrow", "Lionel Messi" and "Elon Musk".

pix2pix. We further compared our method, Control4D, with the baseline approaches including Tensor4D, Tensor4D+GAN, and GaussianPlanes to validate the efficiency of our proposed representation and GAN. We evaluated the Fréchet Inception Distance (FID) metric [18] between the edited dataset and generated images. We also compute CLIP cosine similarity [53] between the generated images and text. Compared with Tensor4D and Tensor4D+GAN, our method achieves superior performance, which demonstrates the efficiency of GaussianPlanes. The results also reveals that our method outperforms the baseline and InstructNeRF2NeRF [17] significantly, demonstrating the effectiveness of our proposed 4D editing pipeline.

7. Conclusions

In conclusion, Control4D is a novel approach for efficient, high-fidelity and temporally consistent editing in dynamic 4D scenes. It utilizes an efficient 4D representation GaussianPlanes and a 2D diffusion-based editor. By utilizing plane-based decomposition to struct Gaussian Splat-

ting, GaussianPlanes ensure both efficiency and robustness for 4D editing. To tackle with the inconsistency caused by diffusion-based editor, Control4D leverages a GAN to generate from the editor, avoiding direct supervision. Experimental results demonstrate Control4D's effectiveness in achieving photo-realistic and consistent 4D editing, surpassing previous approaches in real-world scenarios. It represents a significant advancement in text-based image editing, particularly for dynamic scenes.

Limitations. Due to utilizing a canonical Gaussian point clouds with flow representation, our approach relies on learning flow within the 4D scenes to exhibit simplicity and smoothness. This poses challenges for our method in effectively handling rapid and extensive non-rigid movements. Furthermore, our method is constrained by ControlNet, which limits the granularity of edits to a coarse level. Consequently, it is unable to perform precise expression or action edits. Our method also requires iterative optimizations for the editing process and cannot be accomplished in a single step.

References

- [1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017. 5
- [2] Omer Bar-Tal, Dolev Ofri-Amar, Rafaail Fridman, Yoni Kassten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 707–723. Springer, 2022. 2
- [3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022. 4, 7
- [4] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 130–141, 2023. 3
- [5] Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K Wong. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. *arXiv preprint arXiv:2304.00916*, 2023. 3
- [6] Duygu Ceylan, Chun-Hao Paul Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. *arXiv preprint arXiv:2303.12688*, 2023. 2
- [7] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 2, 5
- [8] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873*, 2023. 1, 3
- [9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 2
- [10] Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B Tenenbaum, and Jiajun Wu. Neural radiance flow for 4d view synthesis and video processing. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14304–14314. IEEE Computer Society, 2021. 3
- [11] Patrick Esser, Johnathan Chiu, Parmida Atighchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. *arXiv preprint arXiv:2302.03011*, 2023. 2
- [12] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12479–12488, 2023. 1, 3
- [13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2, 4
- [14] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenarf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021. 2
- [15] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 2
- [16] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 6
- [17] Ayaan Haque, Matthew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. *arXiv preprint arXiv:2303.12789*, 2023. 1, 2, 3, 8
- [18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 8
- [19] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2
- [20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2
- [21] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2
- [22] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 2
- [23] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 1
- [24] Xin Huang, Ruizhi Shao, Qi Zhang, Hongwen Zhang, Ying Feng, Yebin Liu, and Qing Wang. Humannorm: Learning normal diffusion model for high-quality and realistic 3d human generation. *arXiv preprint arXiv:2310.01406*, 2023. 3
- [25] Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. Humanrf: High-fidelity neural radiance fields for humans in motion. *arXiv preprint arXiv:2305.06356*, 2023. 3
- [26] Dadong Jiang, Zihui Ke, Xiaobo Zhou, and Xidong Shi. 4d-editor: Interactive object-level editing in dynamic neural radiance fields via 4d semantic segmentation. *arXiv preprint arXiv:2310.16858*, 2023. 1
- [27] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)*, 42(4):1–14, 2023. 2, 3

- [28] Levon Khachatryan, Andranik Moysisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023. 2
- [29] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021. 2
- [30] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5521–5531, 2022. 3, 1
- [31] Weiyu Li, Rui Chen, Xuelin Chen, and Ping Tan. Sweet-dreamer: Aligning geometric priors in 2d diffusion for consistent text-to-3d. *arXiv preprint arXiv:2310.02596*, 2023. 2
- [32] Yu-Jhe Li and Kris Kitani. 3d-clfusion: Fast text-to-3d rendering with contrastive latent diffusion. *arXiv preprint arXiv:2303.11938*, 2023. 3
- [33] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. *arXiv preprint arXiv:2211.10440*, 2022. 1, 3
- [34] Haotong Lin, Sida Peng, Zhen Xu, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Efficient neural radiance fields for interactive free-viewpoint video. In *SIGGRAPH Asia Conference Proceedings*, 2022. 1
- [35] Jia-Wei Liu, Yan-Pei Cao, Weijia Mao, Wenqiao Zhang, David Junhao Zhang, Jussi Keppo, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Devrf: Fast deformable voxel radiance fields for dynamic scenes. *arXiv preprint arXiv:2205.15723*, 2022. 1, 3
- [36] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Transactions on Graphics (TOG)*, 40(6):1–16, 2021. 3
- [37] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023. 3
- [38] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. *arXiv preprint arXiv:2308.09713*, 2023. 3
- [39] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. *arXiv preprint arXiv:2211.07600*, 2022. 1
- [40] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2
- [41] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 1
- [42] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7588–7597, 2019. 2
- [43] Thu H Nguyen-Phuoc, Christian Richardt, Long Mai, Yongliang Yang, and Niloy Mitra. Blockgan: Learning 3d object-aware scene representations from unlabelled images. *Advances in neural information processing systems*, 33:6767–6778, 2020. 2
- [44] 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
- [45] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3504–3515, 2020. 3
- [46] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 3
- [47] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021. 1, 3
- [48] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021. 3
- [49] Ryan Po and Gordon Wetzstein. Compositional 3d scene generation using locally conditioned diffusion. *arXiv preprint arXiv:2303.12218*, 2023. 3
- [50] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1, 3
- [51] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 1, 3
- [52] Yi-Ling Qiao, Alexander Gao, and Ming Lin. Neuphysics: Editable neural geometry and physics from monocular videos. *Advances in Neural Information Processing Systems*, 35:12841–12854, 2022. 1
- [53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

- Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 8
- [54] 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
- [55] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 1
- [56] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1
- [57] Hoigi Seo, Hayeon Kim, Gwanghyun Kim, and Se Young Chun. Ditto-nerf: Diffusion-based iterative text to omnidirectional 3d model. *arXiv preprint arXiv:2304.02827*, 2023. 3
- [58] SG_161222. Realistic vision v5 stable diffusion checkpoint, 2023. 1
- [59] Ruizhi Shao, Zerong Zheng, Hanzhang Tu, Boning Liu, Hongwen Zhang, and Yebin Liu. Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16632–16642, 2023. 1, 2, 4
- [60] QiuHong Shen, Xingyi Yang, and XinChao Wang. Anything-3d: Towards single-view anything reconstruction in the wild. *arXiv preprint arXiv:2304.10261*, 2023. 3
- [61] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 3
- [62] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2
- [63] Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, et al. Text-to-4d dynamic scene generation. *arXiv preprint arXiv:2301.11280*, 2023. 3
- [64] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 2
- [65] Liangchen Song, Anpei Chen, Zhong Li, Zhang Chen, Lele Chen, Junsong Yuan, Yi Xu, and Andreas Geiger. Nerf-player: A streamable dynamic scene representation with decomposed neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2732–2742, 2023. 3
- [66] Jingxiang Sun, Bo Zhang, Ruizhi Shao, Lizhen Wang, Wen Liu, Zhenda Xie, and Yebin Liu. Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior. *arXiv preprint arXiv:2310.16818*, 2023. 3
- [67] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 2
- [68] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. *arXiv preprint arXiv:2303.14184*, 2023. 3
- [69] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, Wang Yifan, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, et al. Advances in neural rendering. In *Computer Graphics Forum*, pages 703–735. Wiley Online Library, 2022. 2
- [70] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12959–12970, 2021. 3
- [71] Chaoyang Wang, Ben Eckart, Simon Lucey, and Orazio Gallo. Neural trajectory fields for dynamic novel view synthesis. *arXiv preprint arXiv:2105.05994*, 2021. 3
- [72] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. *arXiv preprint arXiv:2212.00774*, 2022. 3
- [73] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *Advances in Neural Information Processing Systems*, 34:27171–27183, 2021. 3
- [74] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 1
- [75] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 2023. 3
- [76] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humanerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16210–16220, 2022. 3
- [77] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. *arXiv preprint arXiv:2310.08528*, 2023. 3
- [78] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems*, 29, 2016. 2

- [79] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *arXiv preprint arXiv:2212.11565*, 2022. 2
- [80] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9421–9431, 2021. 3
- [81] Zhen Xu, Sida Peng, Haotong Lin, Guangzhao He, Jiaming Sun, Yujun Shen, Hujun Bao, and Xiaowei Zhou. 4k4d: Real-time 4d view synthesis at 4k resolution. *arXiv preprint arXiv:2310.11448*, 2023. 3
- [82] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5746–5756, 2021. 6
- [83] Lvmi Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 3, 7, 1
- [84] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. Structured local radiance fields for human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15893–15903, 2022. 3
- [85] Zerong Zheng, Xiaochen Zhao, Hongwen Zhang, Boning Liu, and Yebin Liu. Avatarrex: Real-time expressive full-body avatars. *arXiv preprint arXiv:2305.04789*, 2023. 3
- [86] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 2
- [87] Jun-Yan Zhu, Zhoutong Zhang, Chengkai Zhang, Jiajun Wu, Antonio Torralba, Josh Tenenbaum, and Bill Freeman. Visual object networks: Image generation with disentangled 3d representations. *Advances in neural information processing systems*, 31, 2018. 2

Control4D: Efficient 4D Portrait Editing with Text

Supplementary Material

A. Implementation Details

A.1. GaussianPlanes

Spatial triplane decomposition. In 3D spatial space, we decompose the attributes of each point in the canonical Gaussian point cloud, including color, opacity, latent feature, and rotation. The decomposition of each attribute utilizes three corresponding feature planes. We employ a HashGrid [41] to represent each feature surface, with a hierarchy of resolutions at 16 levels, where the scale of each level is 1.3 times that of the preceding level. Each level contains 2 feature channels, and the encoded results are mapped to the corresponding attributes through a 256-unit MLP network.

4D flow decomposition. For the 4D flow, we have implemented a hierarchical decomposition using Tensor4D [59]. In our approach, we decompose the flow of each point’s position and rotation at each moment within the Gaussian point cloud. This decomposition employs 9 feature planes. Each feature plane is represented using a HashGrid consistent with the spatial feature planes. Subsequently, the encoded results are first fused individually for the corresponding three planes using three 256-unit MLPs, and then a final attribute output is produced through another 256-unit MLP.

A.2. 4D Generator

Network Structure. We adopt a network architecture similar to pix2pixHD [74] for the GAN’s generator and discriminator. In our generator, we introduce some modifications: the input features comprise both RGB and latent features, which are concatenated together as the input. Additionally, in the intermediate layers, we concatenate the feature with its global feature code. The architecture includes three downsample layers, three middle blocks, and five upsample layers, thereby achieving a 4x super-resolution in the final output. The number of the base feature channel in the network is 32. The input for the 4D generator is at a resolution of 256, and it outputs images at a resolution of 1024. As for the discriminator, we utilize the same architecture as the pix2pixHD discriminator.

Global Encoder. We utilize MobileNet [23] to extract the global code. Initially, the image is resized to a resolution of 224, followed by feature extraction through MobileNet’s layers. We map the final feature of MobileNet to a 64-dimensional global feature code.

Local Encoder. We employ an encoder similar to the VAE encoder used in Stable Diffusion [55] as our local encoder. Our local encoder compresses the original image to a quarter of its original size through two downsample layers, and

the number of “z.channels” is set to 4. The base channel number of our network is 32.

A.3. Diffusion-based Editor

We utilize ControlNet [83] as our diffusion-based editor. To achieve better control effects, we employ both normal and OpenPose as control signals. The control strength for normal is set at 0.5, while for OpenPose, it is 1.0. Additionally, we use the RealisticVision [58], an SD1.5 model, to obtain more realistic editing effects. Additionally, before feeding the images into ControlNet, we resize the 1024-resolution images down to a resolution of 512.

A.4. 4D Reconstruction based on GaussianPlanes

We initially reconstruct the 4D scenes using GaussianPlanes. Our experiments primarily focus on the Tensor4D dataset, and we also showcase some results in challenging scenes, including those from Neural3DVideo [30], ENeRF [34] and InstructNeRF2NeRF [17]. For the Tensor4D dataset, we employ a Gaussian sphere for initialization, with a point cloud size of 5,000 and a radius of 1. For the ENeRF, Neural3DVideo and InstructNeRF2NeRF datasets, we utilize the point cloud from the first frame of COLMAP [56] as the initialization.

During the training process, to ensure the stability of the canonical Gaussian point cloud, we adopt a weighted strategy for selecting training frames. There is a 50% probability that we choose all frames from the first moment and a 50% probability that we randomly select frames from other moments. This approach is designed to balance the representation of the initial frame with the dynamic aspects of the remaining video content. Simultaneously, we are also training the 4D generator in preparation for 4D editing.

During the training process, the learning rate for the point cloud positions linearly decays from 0.00016 to 0.0000016. The learning rates for scaling, color, opacity, and rotation are set at 0.005, while the learning rate for flow is 0.00025. The learning rate for the 4D generator is 0.001. The gradient threshold for splitting is set to 0.0002, and the interval of densification and pruning is 200. We employ L1 loss to train GaussianPlanes, with a weight of 1.0. For training the generator, we use L1 loss, perceptual loss, and GAN loss, with weights of 1.0, 1.0, and 0.01, respectively. The discriminator is trained using GAN loss and gradient penalty regularization, with respective weights of 1.0 and 0.01.



Figure 7. Ablation study of 4D generator. First row: Results utilizing only GaussianPlanes, second row: Results achieved by combining 4D generator. The prompts used here are "Elf King" and "Doctor Strange".



Figure 8. Control4D results on ENeRF dataset. The prompts are "Iron Man" and "Lionel Messi".

A.5. 4D Editing Process

During the 4D editing process, we utilize two GPUs (RTX3090) for training. One GPU is dedicated to running edits for each image, while the other GPU is tasked with running GaussianPlanes and rendering images with the 4D generator. These two processes are executed in parallel. For complex multi-camera 4D scenes, including Neural3DVideo and ENeRF, we do not edit using all cameras. Instead, we use images from all cameras at the first moment

and randomly select images from four cameras at other moments to form the dataset.

The first 1000 steps of our training are for static editing, followed by 4000 steps for dynamic editing. During static editing, the noise added to the diffusion-based editor is $U(0.02, 0.98)$, which is reduced to $U(0.02, 0.6)$ for dynamic editing. The steps of diffusion model is set to 20 and we use the DDIM scheduler. To enhance robustness, we lower the learning rates during the editing process. Specifi-



Figure 9. Ablation Study of GaussianPlanes. First row: w/o spatial triplane decomposition. Second row: w/o 4D flow decomposition. Third row: Control4D results.

cally, the learning rate for point positions is 0.000016, while the learning rates for scaling, color, opacity, and rotation remain at 0.005, and the learning rate for flow is 0.0001. The 4D generator’s learning rate is set at 0.0001. In the editing process, we don’t split or prune Gaussian points. In the multi-level guidance, the probability of selecting each level is equal. The weights of the various losses in 4D editing remain consistent with those in the reconstruction process.

B. More Comparisons

We conducted further comparisons with InstructNeRF2NeRF on their dataset. As shown in Fig. 13, our method noticeably surpasses InstructNeRF2NeRF in terms of realism and quality. Additionally, our optimization process is extremely fast, completing editing tasks in just 5 minutes, whereas InstructNeRF2NeRF requires at least about 5 hours. This makes our method 60 times more

efficient than InstructNeRF2NeRF.

C. More Ablation Study

GaussianPlanes. We conducted more ablation studies on GaussianPlanes. As shown in Fig. 9, when the spatial triplane decomposition is not used, the results exhibit significant noise. Without the decomposition of flow, the edited results become noticeably blurred, with evident occurrences of jittering, which can be clearly observed in the supplementary video. This demonstrates that our proposed plane decomposition method makes Gaussian Splatting more structured and significantly enhances its robustness.

4D generator. More ablation experiments were conducted on our proposed 4D generator. As illustrated in Fig. 7, without the use of the generator and relying solely on GaussianPlanes, the images noticeably lose many high-quality de-



Figure 10. More Control4D results on Tensor4D dataset. The prompts are "Joe Biden wearing suit", "Donald Trump wearing suit", "Trinity in The Matrix", "Neo in The Matrix", "James Gordon in Batman", and "Joker in Batman".



Figure 11. Ablation study of multi-level guidance.

tails and appear blurry. This validates the role of our 4D generator in enhancing quality.

Multi-level guidance. Further ablation studies were performed on multi-level guidance. As shown in Fig. 11, when only GAN loss is used, mode collapse occurs easily due to the small size of the dataset. When only the first and third levels are used, the results become blurred. This indicates that our progressive guidance strategy improves the stability of the GAN and gradually enhances the quality of the rendered images.

D. More Results

Our method is applicable not only to the editing of half-body and heads but also to complex 4D scenes and full-body human editing. More results are illustrated in Fig. 12, 10, and 8. For dynamic editing effects, please refer to our supplementary video.

E. Social Impact

The primary goal of our method is to provide users with an advanced tool for dynamic human editing in complex 4D environments. While our approach enables intricate editing of full-body humans and facilitates creative expression in digital environments, it also raises concerns about potential misuse, such as creating deceptive or misleading content. This challenge is not exclusive to our method but is a common issue across various generative modeling techniques. Additionally, in line with ethical considerations, our approach underscores the importance of diversity, including aspects of gender, race, and cultural representation. It is crucial for ongoing and future research in generative modeling to



Figure 12. Control4D result on neural 3D video dataset. The prompt is "Mark Zuckerberg".

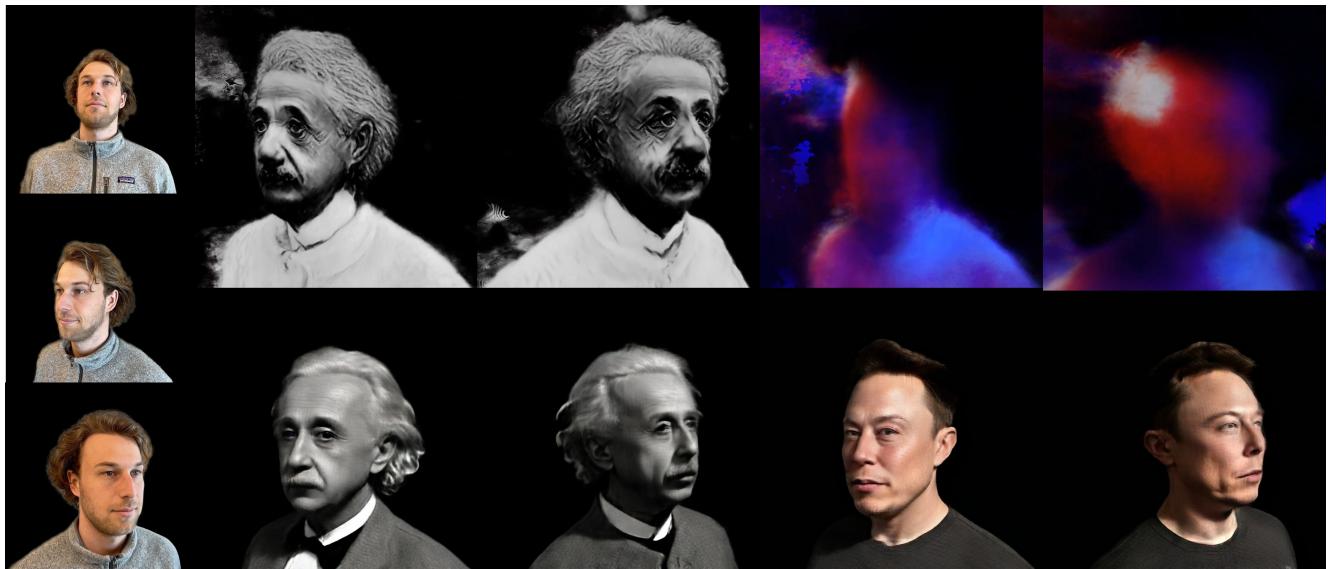


Figure 13. Comparison with InstructNeRF2NeRF on InstructNeRF2NeRF dataset. First row: InstructNeRF2NeRF results with prompts "Turn him into Albert Einstein" and "Turn him into Elon Musk". Second row: Control4D results with prompts "Albert Einstein" and "Elon Musk".

continuously engage with and reevaluate these ethical considerations to ensure responsible use and positive societal impact.