Portrait Video Editing Empowered by Multimodal Generative Priors

XUAN GAO, University of Science and Technology of China, China
HAIYAO XIAO, University of Science and Technology of China, China
CHENGLAI ZHONG, University of Science and Technology of China, China
SHIMIN HU, University of Science and Technology of China, China
YUDONG GUO, University of Science and Technology of China, China
JUYONG ZHANG*, University of Science and Technology of China, China

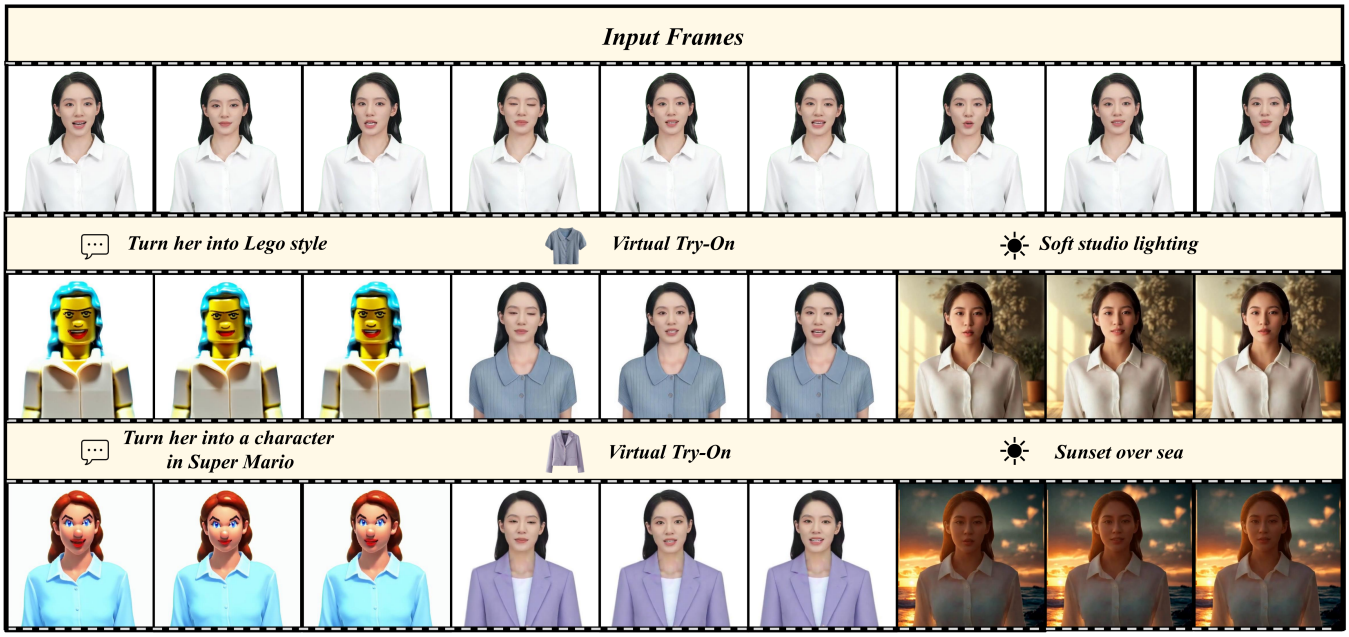

Fig. 1. PortraitGen is a powerful portrait video editing method that achieves consistent and expressive stylization with multimodal prompts. Given a monocular RGB video, our model could perform high-quality text driven editing, image driven editing and relighting.

We introduce PortraitGen, a powerful portrait video editing method that achieves consistent and expressive stylization with multimodal prompts. Traditional portrait video editing methods often struggle with 3D and temporal consistency, and typically lack in rendering quality and efficiency. To address these issues, we lift the portrait video frames to a unified dynamic 3D Gaussian field, which ensures structural and temporal coherence across frames. Furthermore, we design a novel Neural Gaussian Texture mechanism that not only enables sophisticated style editing but also achieves rendering speed over 100FPS. Our approach incorporates multimodal inputs through knowledge distilled from large-scale 2D generative models. Our system also incorporates expression similarity guidance and a face-aware portrait editing module, effectively mitigating degradation issues associated with iterative dataset updates. Extensive experiments demonstrate the temporal consistency, editing efficiency, and superior rendering quality of our method. The broad applicability of the proposed approach is demonstrated through various applications, including text-driven editing, image-driven editing, and relighting, highlighting its great potential to advance the field of video editing. Demo videos and released code are provided in our project page: https://ustc3dv.github.io/PortraitGen/

CCS Concepts: • **Computing methodologies** → **Shape modeling**; *Rendering*; *Machine learning approaches.*

Additional Key Words and Phrases: 4D portrait reconstruction, generative priors, multimodal editing

*Corresponding author (juyong@ustc.edu.cn).

Authors' Contact Information: Xuan Gao, gx2017@mail.ustc.edu.cn, University of Science and Technology of China, China; Haiyao Xiao, xhy1999512@mail.ustc.edu.cn, University of Science and Technology of China, China; Chenglai Zhong, zcl2017@mail.ustc.edu.cn, University of Science and Technology of China, China; Shimin Hu, sa23001018@mail.ustc.edu.cn, University of Science and Technology of China, China; Yudong Guo, yudong@ustc.edu.cn, University of Science and Technology of China, China; Juyong Zhang, juyong@ustc.edu.cn, University of Science and Technology of China, China.

SA Conference Papers '24, December 3–6, 2024, Tokyo, Japan

ACM ISBN 979-8-4007-1131-2/24/12
https://doi.org/10.1145/3680528.3687601

ACM Reference Format:
Xuan Gao, Haiyao Xiao, Chenglai Zhong, Shimin Hu, Yudong Guo, and Juyong Zhang. 2024. Portrait Video Editing Empowered by Multimodal Generative Priors. In *SIGGRAPH Asia 2024 Conference Papers (SA Conference Papers '24), December 3–6, 2024, Tokyo, Japan.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3680528.3687601

1 INTRODUCTION

Portrait video editing has extensive applications in fields such as film, art, and AR/VR. Ensuring structural similarity and temporal consistency across the whole sequence, while enabling various functionalities and modalities, and achieving high-quality editing results, have always been challenging.

2D portrait editing has been studied a lot. Early works [Liu et al. 2022; Yang et al. 2022a,b] mainly adopt Generative Adversarial Network (GAN) [Goodfellow et al. 2014] for editing or stylized animation based on style labels or reference images. By minimizing CLIP [Radford et al. 2021] similarity, some works [Gal et al. 2022; Patashnik et al. 2021; Xia et al. 2021] successfully generate images based on text descriptions. However, this kind of works is limited by the representation ability of the GAN model. Recently, diffusion models [Ho et al. 2020] have shown great generation ability compared with GAN. Based on the denoising diffusion scheme, a lot of generative models, adapters, and finetuning methods are proposed to generate high-quality stylized portrait images. However, when editing portrait videos, these methods struggle to maintain temporal consistency across frames.

To improve the continuity of edited video, some works choose to explore training-free video editing with pre-trained image diffusion models. They use dense correspondence, DDIM inversion [Song et al. 2020], ControlNet [Zhang et al. 2023b], or cross-frame attention to make editing aware of the motion or underlying structures of the original video. Other works turn to connect the frames in temporal dimension and train temporal attention to ensure temporal or multi-view consistency [Guo et al. 2024; Qin et al. 2023]. However, due to the lack of 3D understanding and facial/body priors, they might fail to generate video results that is satisfying in quality and temporal consistency. Meanwhile, these methods need minutes of computation to generate only 1-second video clip due to the progressive sampling and complicated computation of the denoising process.

In this paper, we propose a portrait video editing system that is: (1) preserving portrait structure, (2) temporally consistent, (3) efficient, and (4) capable of multimodal editing requirements. Unlike previous works that focus solely on the 2D domain, we lift the portrait video editing problem into 3D to ensure 3D awareness. Additionally, we distill the multimodal editing knowledge from existing 2D generative models to facilitate high-quality editing.

Specifically, we employ 3D Gaussian Splatting (3DGS) [Kerbl et al. 2023] for consistent and efficient rendering. We embed the 3D Gaussian field on the surface of SMPL-X [Pavlakos et al. 2019] to ensure structural and temporal consistency. Previous 3DGS-based portrait representations [Qian et al. 2023; Xiang et al. 2024] store spherical harmonic (SH) coefficients for each Gaussian and supervise the splatted image directly. However, although these kinds of representations may exhibit high-fidelity in the reconstruction task, they are not qualified for editing tasks. The reason behind this is that many styles include intricate brush strokes and contour lines, which are actually not totally 3D consistent. Directly fitting such signals with 3DGS can result in blurring or artifacts. Moreover, in some artistic styles, portraits often deviate greatly from real people, which calls for more expressive representations. Inspired by Neural Texture [Thies et al. 2019] and screen post-processing effects in non-photorealistic rendering, we store a learnable feature for each Gaussian instead of storing SH coefficients. We then employ a 2D neural renderer to transform the splatted feature map into RGB signals. This approach provides a more informative feature than SH coefficients and allows for a better fusion of splatted features, facilitating the editing of more complex styles. As demonstrated in Fig. 2, with the help of this Neural Gaussian Texture mechanism, our method supports editing whose styles are not completely 3D consistent and achieves rendering speed over 100FPS.

Fig. 2. A totally 3D consistent model may not be an ideal solution for some styles. Many styles include intricate brush strokes and contour lines, which are actually not 3D consistent. Given the instruction 'Turn her into pixel style', our edited portrait could exhibit pixel contour lines, which is crucial for this kind of stylization.

To distill the knowledge of 2D multimodal generative models into portrait video editing, we alternate between editing the dataset of video frames and updating the underlying 3D portrait, inspired by [Haque et al. 2023]. However, we find that naively using this iterative dataset update strategy may accumulate errors in expressions and facial structures, causing blurring and expression degradation. To address these issues, we design an expression similarity guidance term to ensure expression correctness. Additionally, we propose a face-aware portrait editing module to preserve facial structures. Experiments demonstrate that our scheme could effectively preserve personalized structures of original portrait videos and outperform previous works in quality, efficiency, and temporal consistency. Applications such as text-driven editing, image-driven editing, and relighting further underscore the effectiveness and multimodal generalizability of our approach.

In summary, the main contributions of our work include:

- We present PortraitGen, an expressive and consistent portrait video editing system. By lifting the 2D portrait video editing problem into 3D and introducing 3D human priors, it effectively ensures both 3D consistency and temporal consistency of the edited video.
- Our Neural Gaussian Texture mechanism enables richer 3D information and improves the rendering quality of edited portraits, and it helps to support complex styles.

- Our expression similarity guidance and face-aware portrait editing module can effectively handle the degradation problems of iterative dataset update, and further enhance expression quality and preserve personalized facial structures.

2 RELATED WORK

Digital Portrait Representation. Digital portrait representation has been studied for a long time. Blanz and Vetter proposed 3DMM [Blanz and Vetter 1999] to embed 3D head shape into several low-dimensional PCA spaces. The explicit head model has been further studied by a lot of following works. To improve its representation ability, some work extends it to multilinear models [Cao et al. 2013; Vlasic et al. 2006], and non-linear models [Guo et al. 2021; Tran and Liu 2018], articulated models [Li et al. 2017]. They have been used for many applications. However, due to the limited representation ability, they fail to synthesize photo-realistic results.

Implicit representations have been widely used in 3D modeling [Song et al. 2024; Verbin et al. 2022; Wang et al. 2021] and editing [Chong Bao and Bangbang Yang et al. 2022; Qiu et al. 2024; Zhang et al. 2022]. They use neural functions to fit the radiance field, signed distance field, or occupancy field. A series of generative head models have been proposed [Chan et al. 2022, 2021; Deng et al. 2022; Gu et al. 2022; Niemeyer and Geiger 2021; Or-El et al. 2022; Wang et al. 2023b]. Some works proposed parametric implicit head model [Hong et al. 2022; Zhuang et al. 2022] or integrate 3D generative model with face priors [Sun et al. 2023; Wu et al. 2022, 2023b] to realize animation. Although implicit representations could achieve satisfied rendering quality, they suffer from limited rendering efficiency.

Recently, 3D Gaussian Splatting (3DGS) [Kerbl et al. 2023] has been applied to digital head modeling. Because of its flexible representation and fast differentiable rasterizer, this kind of head model achieves remarkable performance in efficiency [Dhamo et al. 2023; Xiang et al. 2024] and fidelity [Qian et al. 2023; Wang et al. 2024; Xu et al. 2023]. There is also work adopting 3DGS for hair modeling and rendering [Luo et al. 2024].

Diffusion Model in Vision. Denoising Diffusion Model [Ho et al. 2020] has showcased great generative ability in vision. Recent works can be classified into 2D image synthesis [Brooks et al. 2023; Rombach et al. 2022; Ruiz et al. 2023; Zhang et al. 2023b] and 3D scene generation [Chen et al. 2023a; Fang et al. 2024; Haque et al. 2023; Liu et al. 2024; Poole et al. 2022; Tang et al. 2023]. While these approaches can generate high-quality results from arbitrary text prompts, they mainly concentrate on generating or editing individual, static tasks and are not intended to directly edit dynamic scenes, especially 2D/3D portrait videos with complex motion.

As a result, some researchers have shifted their focus to video tasks. The main challenge is the consistency between different frames. To solve this problem, some methods [Geyer et al. 2023; Ku et al. 2024; Molad et al. 2023; Qi et al. 2023; Wang et al. 2023a; Wu et al. 2023a; Zhang et al. 2024a] modify the latent space of the diffusion model and introduce cross-frame attention maps to enhance the consistency of the generated results. However, purely modifying in attention space could not enable the model consistent in details. Rerender-A-Video [Yang et al. 2023] and CoDeF [Ouyang et al. 2023] use optical flow to enhance fine-detailed consistency, which suffers from limited optical flow accuracy and struggles to model complex motion.

Portrait Editing. The editing of the appearance and semantic attributes of digital humans has always attracted a lot of attention. Following the success of StyleGAN2 [Karras et al. 2020], many researchers utilize pre-trained GAN model for facial editing or animation [Abdal et al. 2021; Kwon and Ye 2022; Liu et al. 2022; Patashnik et al. 2021; Tzaban et al. 2022; Yang et al. 2022a,b]. However, due to limited generation ability of StyleGAN2, these methods fail to get robust results in complex motion.

To address this issue, researchers utilize 3D representations as geometric proxies to enhance the 3D consistency in editing. Some methods [Aneja et al. 2023; Canfes et al. 2023] directly employ 3DMM (3D Morphable Model) as geometric representation and utilize generative models to generate corresponding UV textures. These methods suffer from the limited representation ability of mesh models and may lack personality in appearance and motion. Recent works [Abdal et al. 2023; Bao et al. 2024; Sun et al. 2023, 2022] utilize NeRF for the purpose of editing, which is not efficient enough for many applications.

Many recent works use diffusion models to perform editing or generation tasks. Among them, 2D image works mainly focus on the generation and editing of face portraits [Papantoniou et al. 2024; Tian et al. 2024]. With the help of Score Distillation Sampling [Poole et al. 2022], researchers tend to construct 3D avatars according to text prompt [Han et al. 2023; Zhang et al. 2023a]. For 3D avatar editing, Avatarstudio [Mendiratta et al. 2023] proposes a view-and-time-aware Score Distillation Sampling to enable high-quality personalized editing across the view and time domain. Control4D [Shao et al. 2024] uses a generative adversarial strategy to handle inconsistency between different frames. Both methods require multi-view dynamic video sequences as input for avatar modeling, which are difficult to obtain for practical use.

3 METHOD

As depicted in Fig. 3, we develop a system that effectively distills knowledge from multimodal generative models to enable consistent, high-quality, and multimodal portrait video editing. To ensure consistency across frames, we propose a 3D portrait representation utilizing 3DGS and holistic human body priors (Sec. 3.2). For high-quality rendering and expressive editing, we incorporate a Neural Gaussian Texture mechanism (Sec. 3.2.1). To support multimodal editing, we introduce specific techniques for text driven editing, image driven editing, and relighting. And we propose strategies to enhance the awareness of expressions and facial structures (Sec. 3.3). In the following, we first provide the preliminary knowledge of the 3DGS and SMPL-X models in Sec. 3.1, and then introduce the technical details.

3.1 Preliminary

3.1.1 3D Gaussian Splatting. 3DGS chooses 3D Gaussians as geometric primitives to represent scenes. Every Gaussian is defined by

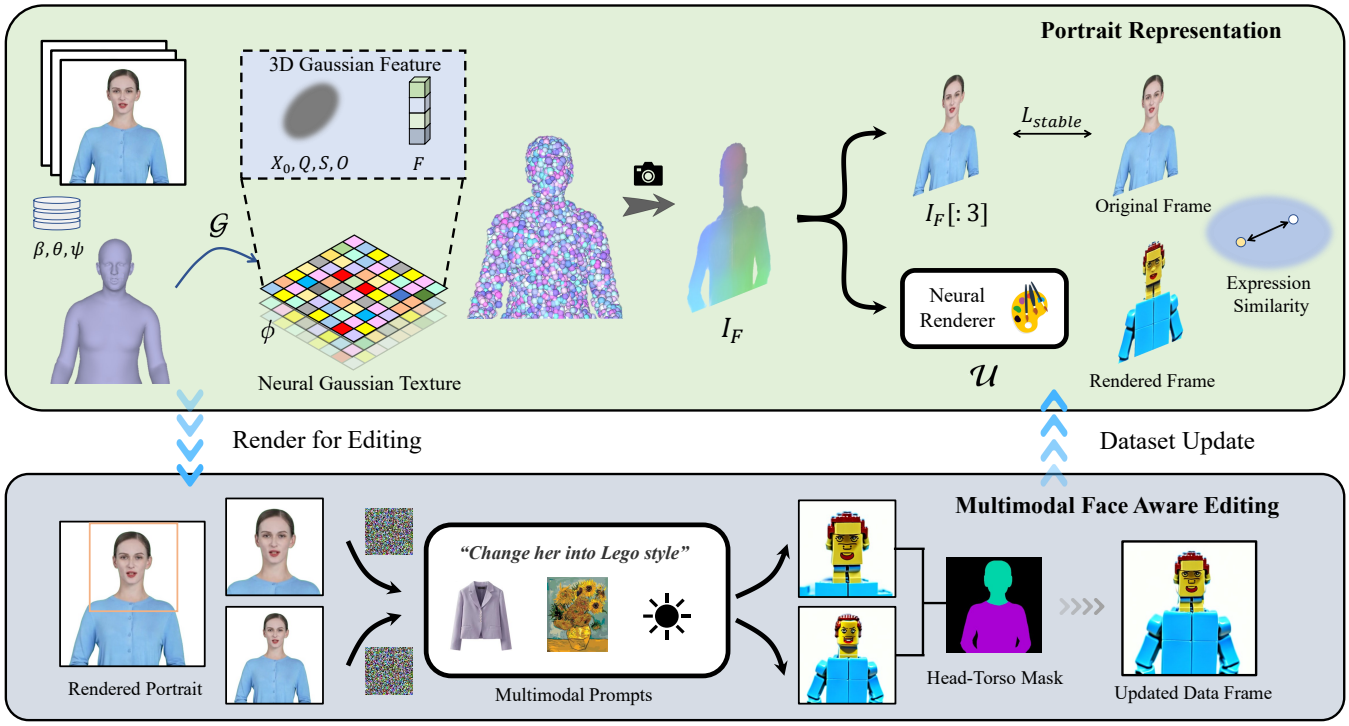

Fig. 3. We first track the SMPL-X coefficients of the given monocular video, and then use a Neural Gaussian Texture mechanism to get a 3D Gaussian feature field. These neural Gaussians are further splatted to render portrait images. An iterative dataset update strategy is applied for portrait editing, and a Multimodal Face Aware Editing module is proposed to enhance expression quality and preserve personalized facial structures.

a 3D covariance matrix Σ centered at point $\mathbf{x_0}$:

$$g(\mathbf{x}) = e^{-\frac{1}{2}(\mathbf{x}-\mathbf{x_0})^T\Sigma^{-1}(\mathbf{x}-\mathbf{x_0})}. \tag{1}$$

Σ is decomposed into a rotation matrix R and a scaling matrix Λ corresponding to learnable quaternion $\mathbf{q}$ and scaling vector $\mathbf{s}$:

$$\Sigma = R\Lambda\Lambda^T R^T. \tag{2}$$

Each 3D Gaussian is attached another two attributes: opacity o and SH coefficients $\mathbf{h}$. The final color for a given pixel is calculated by sorting and blending the overlapped Gaussians:

$$\mathbf{C} = \sum_{i \in N} \mathbf{c}_i \alpha_i \prod_{j=1}^{i-1}(1-\alpha_j), \tag{3}$$

where α_i is computed by the multiplication of projected Gaussian and o. Gaussian field can be denoted as $\{\mathbf{x_0}, \mathbf{q}, \mathbf{s}, o, \mathbf{h}\}$.

3.1.2 SMPL-X. SMPL-X model [Pavlakos et al. 2019] is a holistic, expressive body model, and is defined by a function $M(\beta, \theta, \psi) : \mathbb{R}^{|\beta|\times|\theta|\times|\psi|} \rightarrow \mathbb{R}^{3V}$:

$$M(\beta, \theta, \psi) = W(T(\beta, \theta, \psi), J(\beta), \theta, \mathcal{W}), \tag{4}$$

$$T(\beta, \theta, \psi) = \bar{T} + B_S(\beta; \mathcal{S}) + B_E(\psi; \mathcal{E}) + B_P(\theta; \mathcal{P}). \tag{5}$$

β, θ, ψ are shape, pose and expression parameters, respectively. $B_S(\beta; \mathcal{S})$, $B_P(\theta; \mathcal{P})$, $B_E(\psi; \mathcal{E})$ are the blend shape functions. Blend skinning function $W(\cdot)$ [Lewis et al. 2000] rotates the vertices in $T(\cdot)$ around the estimated joints $J(\beta)$ smoothed by blend weights.

To model long hairs and loose clothing, we introduce a learnable vertices displacement and the final mesh is computed as:

$$\widehat{M}(\beta, \theta, \psi) = M(\beta, \theta, \psi) + \Delta M. \tag{6}$$

3.2 Portrait Representation

To achieve high-fidelity and efficient rendering, we utilize dynamic 3DGS as the portrait avatar representation. Although naively using color or SH coefficient $\mathbf{h}$ may be enough for reconstruction tasks like previous representations [Li et al. 2024; Xiang et al. 2024; Zielonka et al. 2023a], it is not enough for editing task, especially for some complex styles. Many styles are not inherently 3D-consistent, as demonstrated in Fig. 2, directly fitting these signals with a 3D model may introduce blur or artifacts. Some styles also have complex structures, which is hard to be optimized for pure 3D models. To improve the representation ability and make it possible to edit with complex styles, we introduce a novel Neural Gaussian Texture mechanism.

3.2.1 Neural Gaussian Texture. Similar to FlashAvatar [Xiang et al. 2024], we maintain a 3D Gaussian field on the UV space of the SMPL-X model, and further deform the Gaussians according to the deformation of underlying meshes tracked from the input video. By embedding a 3D Gaussian field on the surface, the 3D Gaussian field could be efficiently transformed by parameters β, θ, ψ. Inspired by Neural Texture proposed by Defered Neural Rendering [Thies et al. 2019], we store learnable features for each Gaussian, instead of storing spherical harmonic coefficients. To be specific, we have a Neural

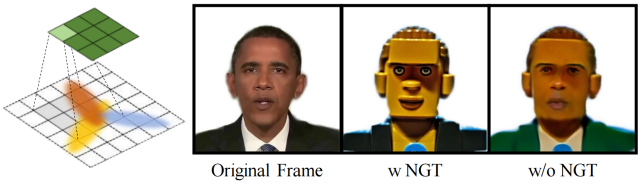

Fig. 4. The Neural Renderer could effectively combine the information of splatted Gaussians and further improve the representation ability of 3D Gaussian portrait representation. With our Neural Gaussian Texture mechanism, the edited portrait follows prompts better and exhibit higher quality. (given instruction: Turn him into Lego style)

Gaussian Field ϕ in the UV field where each pixel is characterized by four attributes: neural feature, opacity, scales, and rotation. Using UV mapping $\mathcal{G} \in \mathbb{R}^3 \rightarrow \mathbb{R}^2$, we transform neural Gaussians from UV space to 3D space. This operation $\mathcal{F}$ could be written as:

$$(X_0, Q, S, O, F) = \mathcal{F}(\widehat{M}(\beta, \theta, \psi), \mathcal{G}, \phi), \tag{7}$$

Given $\widehat{M}(\beta, \theta, \psi)$, $\mathcal{G}$ and ϕ, we could get the embeded 3D Gaussian field (X_0, Q, S, O, F) corresponding to a certain frame.

3.2.2 Neural Rendering. Given camera intrinsic parameters K, camera poses $P = \{P_i\}_{i=1}^{N}$, and the 3D Gaussian field, we perform differentiable tile renderer $\mathcal{R}$ to render a feature image. Then the feature image is operated by a 2D Neural Renderer $\mathcal{U}$ to convert it to RGB domain:

$$I_F = \mathcal{R}((X_0, Q, S, O, F), K, P), \tag{8}$$

$$I = \mathcal{U}(I_F). \tag{9}$$

I and I_F share the same resolution, and they are all 512×512 in our setting.

Many styles differ greatly from real people or are not totally 3D consistent. As shown in Fig. 4, our 2D Neural Renderer operates on the splatted feature map. Our Neural Gaussian Texture mechanism improves the model's capacity and could effectively combine the information of splatted Gaussians, which further improves the representation ability.

3.2.3 Reconstruction Details. We reconstruct the personalized 3D Gaussian Avatar with the following loss terms:

Reconstruction Loss. This loss requires that the rendered result is consistent with the input RGB image, which is common for RGB reconstruction and can be formulated as:

$$L_{recon}(I, I_{src}) = \|I - I_{src}\|_1. \tag{10}$$

Mask Loss. This loss requires that the rendered alpha channel A is consistent with the segmentation map of the input source image:

$$L_{mask}(A, A_{src}) = \|A - A_{src}\|_1. \tag{11}$$

Perceptual Loss. The perceptual loss L_{LPIPS} of [Zhang et al. 2018] is utilized to provide robustness to slight misalignments and shading variations and improve details in the reconstruction. We choose VGG as the backbone of LPIPS.

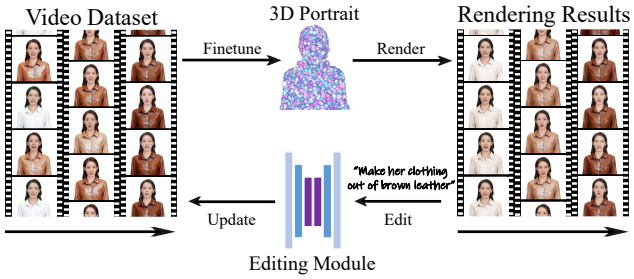

Fig. 5. We alternate between editing the dataset of video frames and updating the underlying 3D portrait. The portrait model will gradually converge to the target prompt, achieving both 3D and temporal consistency.

Stable Loss. We found that training with above three loss terms may be unstable, and we further supervise part of the latent feature space F directly:

$$L_{stable}(I_F, I_{src}) = \|I_F[:3] - I_{src}\|_1, \tag{12}$$

where the first 3 channels of I_F are supervised by input source frames.

In summary, the overall loss of training our model is defined as:

$$\begin{aligned} L_{total} =& \lambda_1 L_{recon}(I, I_{src}) + \lambda_2 L_{mask}(A, A_{src}) \\ &+ \lambda_3 L_{LPIPS}(I, I_{src}) + \lambda_4 L_{stable}(I_F, I_{src}). \end{aligned} \tag{13}$$

3.3 Editing

We employ a variety of pre-trained generative models for multimodal prompt-guided editing. To tackle the issue of inconsistent edits across different frames, as illustrated in Fig. 5, we alternate between editing the dataset of video frames and updating the underlying 3D portrait. Specifically, this process is to repeat as follows: (1) A portrait image is rendered from a training viewpoint. (2) The image is edited by the editing module. (3) The training dataset image is replaced with the edited image. (4) The portrait representation continues training with the updated dataset. The portrait model will gradually converge to the targeting prompt, achieving both 3D and temporal consistency.

To handle degradation problems in expressions and facial structures, we propose an expression similarity guidance term and a face-aware portrait editing module to emphasize facial information.

3.3.1 Expression Similarity Guidance. Although many 2D editing models are claimed to be structure-preserving, they are not very robust to complex expression details. Accumulated errors after many times of editing may further misguide the expressions far from the original video. To enhance expression cognition, we map the rendered image and input source image to the latent expression space of EMOCA [Filntisis et al. 2022], and use a loss function to ensure similarity:

$$L_{exp}(I, I_{src}) = \|\mathcal{E}_{exp}(I) - \mathcal{E}_{exp}(I_{src})\|_2^2. \tag{14}$$

3.3.2 Face-Aware Portrait Editing. When editing an upper body image where the face occupies a relatively small portion, the editing may not be robust enough to detailed facial structure. We further propose a training-free strategy that improves the editing quality of the face region. As shown in Fig. 3, we first crop and resize the

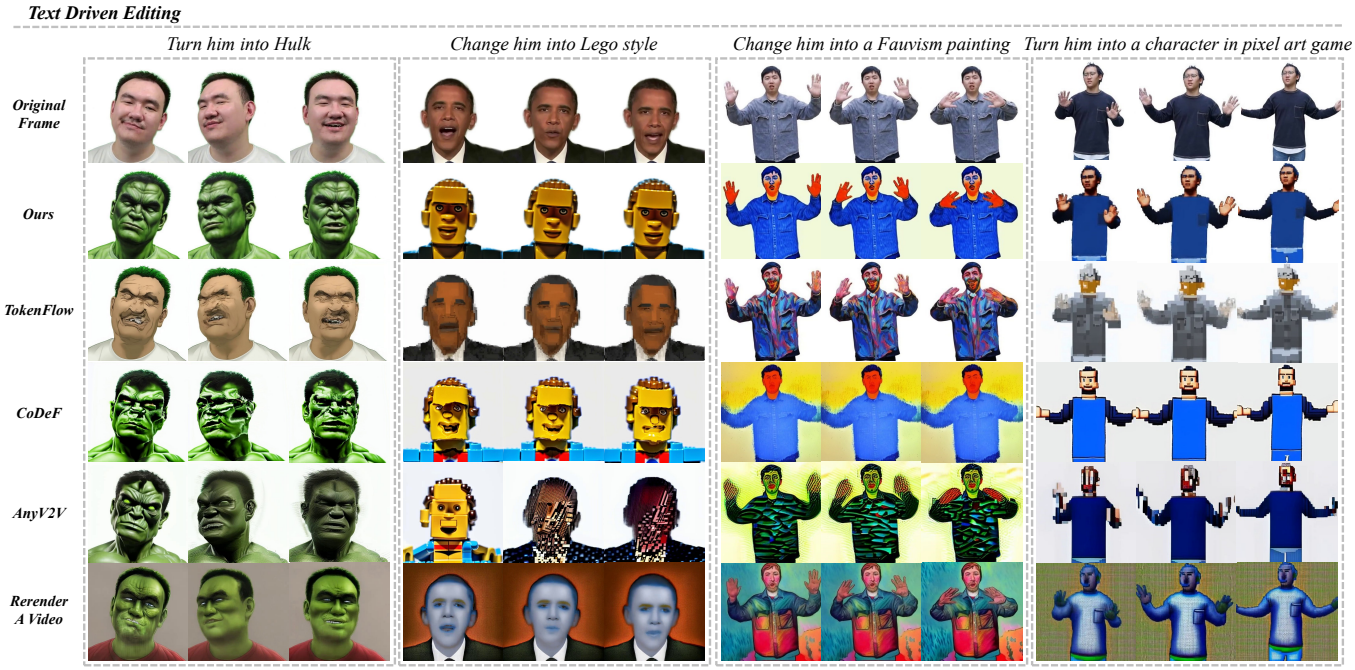

Fig. 6. Qualitative comparisons on text driven portrait editing.

face region into 512×512. Both facial part and portrait part are then edited by the image editing model, and both edited parts are then composited into the final frame image with the head-torso mask.

3.3.3 Editing Details. For each optimization step, we randomly select a frame and use the corresponding SMPL-X parameters to render image I. The selected updated dataset frame is denoted as I^*. The corresponding original image (which is the image that is not edited) is denoted as I_{src}. We finetune the avatar reconstructed in section 3.2 with the following loss function:

$$\begin{aligned} L_{edit} =& \lambda_1 L_{recon}(I, I^*) + \lambda_2 L_{mask}(A, A_{src}) \\ &+ \lambda_3 L_{LPIPS}(I, I^*) + \lambda_4 L_{stable}(I_F, I_{src}) + \lambda_5 L_{exp}(I, I_{src}). \end{aligned} \tag{15}$$

3.4 Applications

Our scheme is a unified portrait video editing framework. Any structure-preserving image editing model could be used to synthesize a 3D consistent and temporally coherent portrait video. In this paper, we demonstrate its effectiveness via several challenging tasks:

3.4.1 Text Driven Editing. We use InstructPix2Pix [Brooks et al. 2023] as a 2D editing model. We add partial noise to the rendered image and edit it based on input source image I_{src} and instruction.

3.4.2 Image Driven Editing. We focus on two kinds of editing works based on image prompts. One kind is to extract the global style of a reference image and another aims to customize an image by placing an object at a specific location. These approaches are utilized in our experiments for style transfer and virtual try-on. We use the method of [Gatys et al. 2016] to transfer the style of a reference image to the dataset frames and use AnyDoor [Chen et al. 2023b] to change the clothes of the subject.

3.4.3 Relighting. We utilize IC-Light [Zhang et al. 2024b] to manipulate the illumination of the video frames. Given a text description as the light condition, our method can harmoniously adjust the lighting of the portrait video.

4 EXPERIMENTS

4.1 Implementation Details

We use the videos released by NeRFBlendshape [Gao et al. 2022], Neural Head Avatar [Grassal et al. 2022], INSTA [Zielonka et al. 2023b] and PointAvatar [Zheng et al. 2023] for validation. Since the released videos only contain the head region, we also collected some datasets from the Internet and captured some monocular videos of the upper body. We use FaRL [Zheng et al. 2022] to get head-torso masks. We use an algorithm similar to TalkSHOW [Yi et al. 2023] for fitting SMPL-X parameters to video frames. It takes about 10 minutes for reconstruction and about 20 minutes for editing. We run our experiments on one RTX 3090 GPU.

4.2 Qualitative Comparison

We compare our method with state-of-the-art video editing methods, including TokenFlow [Geyer et al. 2023], Rerender-A-Video (denoted as RAV) [Yang et al. 2023], CoDeF [Ouyang et al. 2023] and AnyV2V [Ku et al. 2024]. TokenFlow and Rerender-A-Video only support text-driven editing tasks, while CoDeF and AnyV2V could support editing with all modalities. For CoDeF, we train the deformation field and canonical image on the input video first. Then,

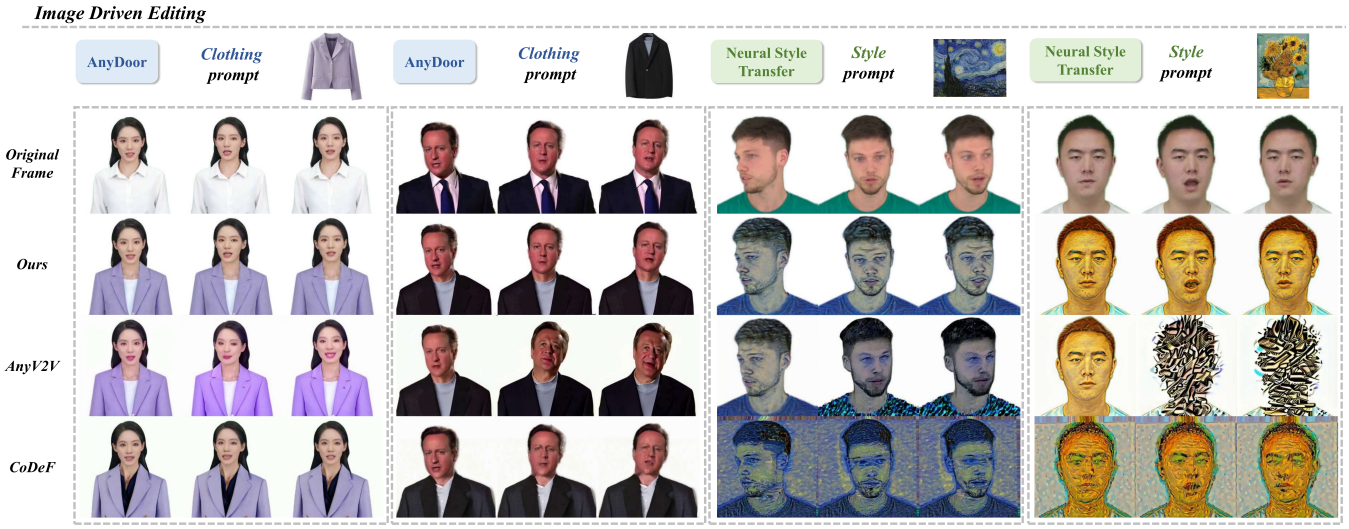

Fig. 7. Qualitative comparisons on image driven portrait editing.

Fig. 8. Qualitative comparisons on relighting.

we edit the canonical image and generate the final edited video according to the deformation field. For AnyV2V, we edit the first frame and then perform image-to-video reconstruction.

To ensure a fair comparison, we limit the video segments used in our evaluation to 2 seconds, each consisting of 60 frames. This is necessary because TokenFlow requires significant GPU memory as the number of frames increases, and CoDeF must learn the deformation fields for the entire video sequence, making it unsuitable for long videos. Although our method can handle videos of arbitrary length, selecting shorter segments allows for a fair evaluation across different methods.

We present qualitative comparisons on text-driven editing in Fig. 6, image-driven editing in Fig. 7, and relighting in Fig. 8. For TokenFlow and Rerender-A-Video, we observe that sometimes the expressions in the edited frames do not maintain consistency with the original video, and the edits fail to align with the given prompts. This may be because extended attention mechanisms can cause the latent codes to drift out of the domain, thereby degrading the quality of the edited results. Additionally, both methods frequently produce noticeable artifacts in the facial regions. This discrepancy can be attributed to the limitations of extended attention in maintaining detailed consistency, especially in capturing facial expressions. Inaccurate correspondences in nearest neighbor search or optical flow estimation further exacerbate these discrepancies.

Although CoDeF's unique modeling approach enhances its capacity to preserve detailed consistency in short video segments,

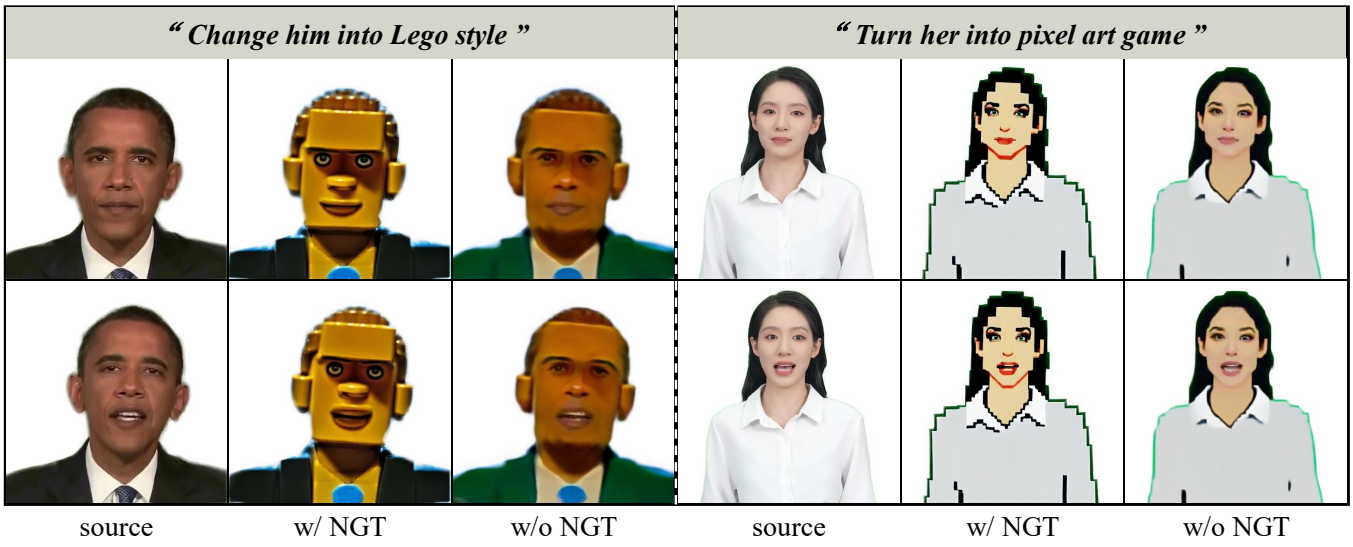

Fig. 9. Neural Gaussian Texture mechanism could remarkably improve the editing results and make it possible to edit with more complex styles.

it fails to generate reasonable results when faced with exaggerated expression and pose changes. This issue primarily stems from the limitations of its 2D deformation field, which is inadequate for modeling complex 3D portrait deformations. We also observe that AnyV2V lacks stability in portrait editing. It frequently fails to maintain consistent appearance and structural integrity, likely due to its unstable editing scheme.

In contrast, our approach leverages a 3DGS-based portrait as the geometric representation, which ensures superior 3D consistency. By integrating prior information about the portrait, we precisely capture changes in expressions and postures, thereby maintaining temporal consistency in the edited results. Moreover, our model adeptly handles challenging multimodal prompts, which can be problematic for other methods. For a more detailed comparison, we encourage viewing the accompanying video.

4.3 Quantitative Comparison

We conducted a user study to further quantitatively validate our method. Participants were asked to watch rendered videos side by side from various methods and respond to a series of questions comparing the results. For each group of editing results, participants addressed the following queries:

- Q1: Which method best follows the given input prompt? (Prompt Preservation)
- Q2: Which method best retains the identity of the input sequence in the video? (Identity Preservation)
- Q3: Which method best maintains temporal consistency in the video? (Temporal Consistency)
- Q4: Which method best preserves expressions and body movements of the input sequence in the video? (Human Motion Preservation)
- Q5: Which method is best overall considering the above four aspects? (Overall)

We collected statistics from 96 participants across 23 groups of editing results. For each case, the video results were randomly shuffled for fair comparison. As shown in Table. 1, our method remarkably outperforms other methods in prompt preservation, identity preservation, temporal consistency, and human motion preservation and is rated as the best in overall quality. These results demonstrate that our approach is highly favored by users, highlighting its effectiveness in various editing dimensions.

	TokenFlow	CoDeF	AnyV2V	RAV	Ours
Q1	8.0	19.1	3.3	3.4	**66.2**
Q2	6.8	7.1	2.6	6.9	**76.6**
Q3	3.9	6.1	1.2	3.5	**85.3**
Q4	3.8	5.1	1.4	4.3	**85.4**
Q5	4.5	6.7	1.4	2.2	**85.2**

Table 1. The table reports the percentages at which a method was rated the best with respect to a specific question. Our method remarkably outperforms other methods in all questions, which demonstrates that our approach is much more likely to be favored by users.

4.4 Editing Efficiency

We further validated the editing efficiency of our method by analyzing the number of frames processed per minute, as illustrated in Table. 2. Different methods employ different modes of inference. For a fair comparison, we compute the time cost including both reconstruction and editing, and average it across the processed frames. We can see that our method outperforms previous video editing methods in terms of efficiency, which further showcases its promising application prospects.

Fig. 10. Expression Similarity Guidance could effectively solve expression degradation problems and keep the expressions consistent with original video frames. (prompt: Change her into a bronze statue.)

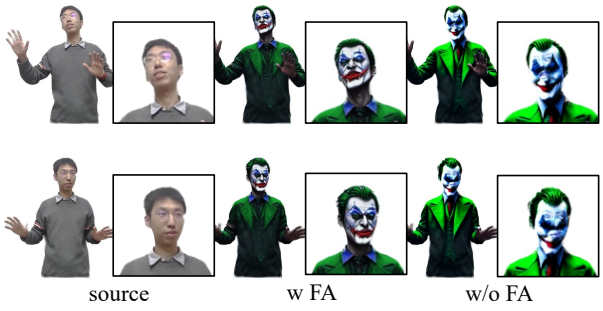

Fig. 11. Naively editing the whole portrait image may cause misalignment of head pose, and blur in the facial region.

TokenFlow	CoDeF	AnyV2V	RAV	Ours
1.6	5.0	1.6	10.0	**60.0**

Table 2. Comparison in editing efficiency. The values represent the number of frames edited per minute. Our method outperforms previous video editing methods, further verifying its promising application prospects.

5 ABLATION STUDY

5.1 Neural Gaussian Texture

Previous portrait representations [Qian et al. 2023; Xiang et al. 2024] use explicit 3D Gaussian for rendering by storing spherical harmonic coefficients for each Gaussian to directly render portrait image. We demonstrate that this approach is unable to represent complex styles like contour lines and brush strokes as they adopt pure 3D representations.

Fig. 9 shows the comparison results between using our Neural Gaussian Texture (NGT) and explicit 3D Gaussian. For the prompt "Change him into Lego style", our method adeptly transforms the editing into the desired shape, while explicit 3D Gaussians struggle to achieve a Lego-like deformation. This is because our Neural Renderer could fuse the features of splatted Gaussians, and further improve its representation ability. For the prompt "Turn her into pixel art game", explicit 3D Gaussians fail to represent contour lines and pixel style elements, demonstrating the limitations of using a purely 3D consistent representation for such stylized edits.

5.2 Face-Aware Portrait Editing

When editing an upper body image where the face occupies a relatively small portion, the model's editing may not be robust enough to head pose and facial structure. Face-Aware Portrait Editing (FA) could enhance the awareness of face structures by performing editing twice. As demonstrated in Fig. 11, naively editing the whole portrait image may cause misalignment of head pose, and blur in the facial region.

5.3 Expression Similarity Guidance

By mapping the rendered image and input source image into the latent expression space of EMOCA, and optimizing for expression similarity, we can further keep the expressions natural and consistent with the original input video frames. As demonstrated in Fig. 10, omitting Expression Similarity Guidance during training leads to expression degeneration.

6 CONCLUSION & DISCUSSIONS

We proposed an expressive multimodal portrait video editing scheme. In contrast to previous approaches that primarily focus on the 2D domain, we elevated the portrait video editing challenge to a 3D perspective. Our method embedded a 3D Gaussian field onto the surface of SMPL-X, ensuring consistency in human body structures across both spatial and temporal domains. Additionally, the proposed Neural Gaussian Texture mechanism could effectively deal with complex styles and achieve rendering speeds of over 100FPS. We leveraged the multimodal editing knowledge of 2D generative models to enhance the quality of 3D editing. Our expression similarity guidance and face-aware portrait editing module effectively handled the degradation problems of iterative dataset updates.

Although we have achieved remarkable improvement in quality and efficiency compared with existing works, there still remain some limitations. Our method relies on tracked SMPL-X, and thus large errors in tracking may cause artifacts. As our method utilizes pretrained 2D editing models for dataset update, the editing ability of our method is restricted by these models. We believe more powerful 2D editing models will further unleash the potential of our paradigm.

ACKNOWLEDGMENTS

This research was supported by the National Natural Science Foundation of China (No.62122071, No.62272433, No.62402468), the Fundamental Research Funds for the Central Universities (No. WK3470000021), and the advanced computing resources provided by the Supercomputing Center of University of Science and Technology of China.

REFERENCES

Rameen Abdal, Hsin-Ying Lee, Peihao Zhu, Menglei Chai, Aliaksandr Siarohin, Peter Wonka, and Sergey Tulyakov. 2023. 3davatargan: Bridging domains for personalized editable avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4552–4562.

Rameen Abdal, Peihao Zhu, Niloy J. Mitra, and Peter Wonka. 2021. StyleFlow: Attribute-conditioned Exploration of StyleGAN-Generated Images using Conditional Continuous Normalizing Flows. *ACM Transactions on Graphics* 40, 3 (may 2021), 1–21. https://doi.org/10.1145/3447648

Shivangi Aneja, Justus Thies, Angela Dai, and Matthias Nießner. 2023. Clipface: Text-guided editing of textured 3d morphable models. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–11.

Chong Bao, Yinda Zhang, Yuan Li, Xiyu Zhang, Bangbang Yang, Hujun Bao, Marc Pollefeys, Guofeng Zhang, and Zhaopeng Cui. 2024. GeneAvatar: Generic Expression-Aware Volumetric Head Avatar Editing from a Single Image. In *The IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*.

Volker Blanz and Thomas Vetter. 1999. A Morphable Model for the Synthesis of 3D Faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*. 187–194.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18392–18402.

Zehranaz Canfes, M Furkan Atasoy, Alara Dirik, and Pinar Yanardag. 2023. Text and image guided 3d avatar generation and manipulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 4421–4431.

Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. 2013. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics* 20, 3 (2013), 413–425.

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. 2022. Efficient geometry-aware 3D generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16123–16133.

Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. 2021. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*. 5799–5809.

Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. 2023b. Anydoor: Zero-shot object-level image customization. *arXiv preprint arXiv:2307.09481* (2023).

Yiwen Chen, Zilong Chen, Chi Zhang, Feng Wang, Xiaofeng Yang, Yikai Wang, Zhongang Cai, Lei Yang, Huaping Liu, and Guosheng Lin. 2023a. GaussianEditor: Swift and Controllable 3D Editing with Gaussian Splatting. arXiv:2311.14521 [cs.CV]

Chong Bao and Bangbang Yang, Zeng Junyi, Bao Hujun, Zhang Yinda, Cui Zhaopeng, and Zhang Guofeng. 2022. NeuMesh: Learning Disentangled Neural Mesh-based Implicit Field for Geometry and Texture Editing. In *European Conference on Computer Vision (ECCV)*.

Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. 2022. GRAM: Generative Radiance Manifolds for 3D-Aware Image Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Helisa Dhamo, Yinyu Nie, Arthur Moreau, Jifei Song, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. 2023. Headgas: Real-time animatable head avatars via 3d gaussian splatting. *arXiv preprint arXiv:2312.02902* (2023).

Jiemin Fang, Junjie Wang, Xiaopeng Zhang, Lingxi Xie, and Qi Tian. 2024. GaussianEditor: Editing 3D Gaussians Delicately with Text Instructions. In *CVPR*.

Panagiotis P. Filntisis, George Retsinas, Foivos Paraperas-Papantoniou, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. 2022. Visual Speech-Aware Perceptual 3D Facial Expression Reconstruction from Videos. *arXiv preprint arXiv:2207.11094* (2022).

Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. StyleGAN-NADA: CLIP-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–13.

Xuan Gao, Chenglai Zhong, Jun Xiang, Yang Hong, Yudong Guo, and Juyong Zhang. 2022. Reconstructing Personalized Semantic Facial NeRF Models From Monocular Video. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)* 41, 6 (2022). https://doi.org/10.1145/3550454.3555501

Leon Gatys, Alexander Ecker, and Matthias Bethge. 2016. A Neural Algorithm of Artistic Style. *Journal of Vision* 16, 12 (2016), 326–326.

Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. 2023. TokenFlow: Consistent Diffusion Features for Consistent Video Editing. *arXiv preprint arxiv:2307.10373* (2023).

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, Vol. 27.

Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. 2022. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18653–18664.

Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. 2022. StyleNeRF: A Style-based 3D Aware Generator for High-resolution Image Synthesis. In *International Conference on Learning Representations*.

Yudong Guo, Lin Cai, and Juyong Zhang. 2021. 3D Face From X: Learning Face Shape From Diverse Sources. *IEEE Trans. Image Process.* 30 (2021), 3815–3827.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. 2024. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. *International Conference on Learning Representations* (2024).

Xiao Han, Yukang Cao, Kai Han, Xiatian Zhu, Jiankang Deng, Yi-Zhe Song, Tao Xiang, and Kwan-Yee K. Wong. 2023. HeadSculpt: Crafting 3D Head Avatars with Text. *arXiv preprint arXiv:2306.03038* (2023).

Ayaan Haque, Matthew Tancik, Alexei Efros, Aleksander Holynski, and Angjoo Kanazawa. 2023. Instruct-NeRF2NeRF: Editing 3D Scenes with Instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html

Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. 2022. HeadNeRF: A Real-time NeRF-based Parametric Head Model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8110–8119.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics* 42, 4 (July 2023). https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/

Max Ku, Cong Wei, Weiming Ren, Harry Yang, and Wenhu Chen. 2024. AnyV2V: A Plug-and-Play Framework For Any Video-to-Video Editing Tasks. *arXiv preprint arXiv:2403.14468* (2024).

Gihyun Kwon and Jong Chul Ye. 2022. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18062–18071.

J. P. Lewis, Matt Cordner, and Nickson Fong. 2000. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 165–172.

Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.* 36, 6 (2017), 194–1.

Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. 2024. Animatable Gaussians: Learning Pose-dependent Gaussian Maps for High-fidelity Human Avatar Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xiangyue Liu, Han Xue, Kunming Luo, Ping Tan, and Li Yi. 2024. GenN2N: Generative NeRF2NeRF Translation. *arXiv preprint arXiv:2404.02788* (2024).

Yuchen Liu, Zhixin Shu, Yijun Li, Zhe Lin, Richard Zhang, and SY Kung. 2022. 3d-fm gan: Towards 3d-controllable face manipulation. In *European Conference on Computer Vision*. Springer, 107–125.

Haimin Luo, Min Ouyang, Zijun Zhao, Suyi Jiang, Longwen Zhang, Qixuan Zhang, Wei Yang, Lan Xu, and Jingyi Yu. 2024. GaussianHair: Hair Modeling and Rendering with Light-aware Gaussians. *arXiv preprint arXiv:2402.10483* (2024).

Mohit Mendiratta, Xingang Pan, Mohamed Elgharib, Kartik Teotia, Mallikarjun B R, Ayush Tewari, Vladislav Golyanik, Adam Kortylewski, and Christian Theobalt. 2023. AvatarStudio: Text-driven Editing of 3D Dynamic Human Head Avatars. arXiv:2306.00547 [cs.CV]

Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. 2023. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329* (2023).

Michael Niemeyer and Andreas Geiger. 2021. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11453–11464.

Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. 2022. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13503–13513.

Hao Ouyang, Qiuyu Wang, Yuxi Xiao, Qingyan Bai, Juntao Zhang, Kecheng Zheng, Xiaowei Zhou, Qifeng Chen, and Yujun Shen. 2023. CoDeF: Content Deformation Fields for Temporally Consistent Video Processing. arXiv:2308.07926 [cs.CV]

Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, Jiankang Deng, Bernhard Kainz, and Stefanos Zafeiriou. 2024. Arc2Face: A Foundation Model of Human Faces. arXiv:2403.11641 [cs.CV]

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2085–2094.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 10975–10985.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. 2022. DreamFusion: Text-to-3D using 2D Diffusion. *arXiv* (2022).

Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. 2023. FateZero: Fusing Attentions for Zero-shot Text-based Video Editing. arXiv:2303.09535 [cs.CV]

Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. 2023. GaussianAvatars: Photorealistic Head Avatars with Rigged 3D Gaussians. *arXiv preprint arXiv:2312.02069* (2023).

Bosheng Qin, Juncheng Li, Siliang Tang, Tat-Seng Chua, and Yueting Zhuang. 2023. InstructVid2Vid: Controllable Video Editing with Natural Language Instructions. arXiv:2305.12328 [cs.CV]

Zherui Qiu, Chenqu Ren, Kaiwen Song, Xiaoyi Zeng, Leyuan Yang, and Juyong Zhang. 2024. Deformable NeRF using Recursively Subdivided Tetrahedra. In *ACM Multimedia 2024*. https://openreview.net/forum?id=QayT1wjqYB

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. http://proceedings.mlr.press/v139/radford21a.html

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22500–22510.

Ruizhi Shao, Jingxiang Sun, Cheng Peng, Zerong Zheng, Boyao Zhou, Hongwen Zhang, and Yebin Liu. 2024. Control4D: Efficient 4D Portrait Editing with Text. (2024).

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Kaiwen Song, Xiaoyi Zeng, Chenqu Ren, and Juyong Zhang. 2024. City-on-Web: Real-time Neural Rendering of Large-scale Scenes on the Web. In *European Conference on Computer Vision (ECCV)*.

Jingxiang Sun, Xuan Wang, Lizhen Wang, Xiaoyu Li, Yong Zhang, Hongwen Zhang, and Yebin Liu. 2023. Next3d: Generative neural texture rasterization for 3d-aware head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20991–21002.

Jingxiang Sun, Xuan Wang, Yong Zhang, Xiaoyu Li, Qi Zhang, Yebin Liu, and Jue Wang. 2022. Fenerf: Face editing in neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7672–7682.

Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. 2023. DreamGaussian: Generative Gaussian Splatting for Efficient 3D Content Creation. *arXiv preprint arXiv:2309.16653* (2023).

Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2019. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)* 38, 4 (2019), 1–12.

Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. 2024. EMO: Emote Portrait Alive - Generating Expressive Portrait Videos with Audio2Video Diffusion Model under Weak Conditions. arXiv:2402.17485 [cs.CV]

Luan Tran and Xiaoming Liu. 2018. Nonlinear 3d face morphable model. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7346–7355.

Rotem Tzaban, Ron Mokady, Rinon Gal, Amit Bermano, and Daniel Cohen-Or. 2022. Stitch it in time: Gan-based facial editing of real videos. In *SIGGRAPH Asia 2022 Conference Papers*. 1–9.

Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. 2022. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 5481–5490.

Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popovic. 2006. Face transfer with multilinear models. In *ACM SIGGRAPH 2006 Courses*. 24–es.

Jie Wang, Jiu-Cheng Xie, Xianyan Li, Feng Xu, Chi-Man Pun, and Hao Gao. 2024. GaussianHead: High-fidelity Head Avatars with Learnable Gaussian Derivation. arXiv:2312.01632 [cs.CV]

Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. 2021. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. *arXiv preprint arXiv:2106.10689* (2021).

Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. 2023b. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4563–4573.

Wen Wang, Kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. 2023a. Zero-Shot Video Editing Using Off-The-Shelf Image Diffusion Models. arXiv:2303.17599 [cs.CV]

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2023a. 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*. 7623–7633.

Yue Wu, Yu Deng, Jiaolong Yang, Fangyun Wei, Chen Qifeng, and Xin Tong. 2022. AniFaceGAN: Animatable 3D-Aware Face Image Generation for Video Avatars. In *Advances in Neural Information Processing Systems*.

Yue Wu, Sicheng Xu, Jianfeng Xiang, Fangyun Wei, Qifeng Chen, Jiaolong Yang, and Xin Tong. 2023b. AniPortraitGAN: Animatable 3D Portrait Generation from 2D Image Collections. In *SIGGRAPH Asia 2023 Conference Proceedings*.

Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. 2021. TediGAN: Text-Guided Diverse Face Image Generation and Manipulation. arXiv:2012.03308 [cs.CV]

Jun Xiang, Xuan Gao, Yudong Guo, and Juyong Zhang. 2024. FlashAvatar: High-fidelity Head Avatar with Efficient Gaussian Embedding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. 2023. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. *arXiv preprint arXiv:2312.03029* (2023).

Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. 2022a. Pastiche master: Exemplar-based high-resolution portrait style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7693–7702.

Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. 2022b. VToonify: Controllable High-Resolution Portrait Video Style Transfer. *ACM Transactions on Graphics (TOG)* 41, 6, Article 203 (2022), 15 pages. https://doi.org/10.1145/3550454.3555437

Shuai Yang, Yifan Zhou, Ziwei Liu, , and Chen Change Loy. 2023. Rerender A Video: Zero-Shot Text-Guided Video-to-Video Translation. In *ACM SIGGRAPH Asia Conference Proceedings*.

Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. 2023. Generating Holistic 3D Human Motion from Speech. In *CVPR*.

Kai Zhang, Nick Kolkin, Sai Bi, Fujun Luan, Zexiang Xu, Eli Shechtman, and Noah Snavely. 2022. ARF: Artistic Radiance Fields.

Longwen Zhang, Qiwei Qiu, Hongyang Lin, Qixuan Zhang, Cheng Shi, Wei Yang, Ye Shi, Sibei Yang, Lan Xu, and Jingyi Yu. 2023a. DreamFace: Progressive Generation of Animatable 3D Faces under Text Guidance. *ACM Trans. Graph.* 42, 4 (2023), 138:1–138:16. https://doi.org/10.1145/3592094

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023b. Adding Conditional Control to Text-to-Image Diffusion Models.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2024b. IC-Light GitHub Page.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

Zicheng Zhang, Bonan Li, Xuecheng Nie, Congying Han, Tiande Guo, and Luoqi Liu. 2024a. Towards consistent video editing with text-to-image diffusion models. *Advances in Neural Information Processing Systems* 36 (2024).

Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. 2022. General Facial Representation Learning in a Visual-Linguistic Manner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18697–18709.

Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J Black, and Otmar Hilliges. 2023. Pointavatar: Deformable point-based head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21057–21067.

Yiyu Zhuang, Hao Zhu, Xusen Sun, and Xun Cao. 2022. Mofanerf: Morphable facial neural radiance field. In *European Conference on Computer Vision*. Springer, 268–285.

Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. 2023a. Drivable 3D Gaussian Avatars. (2023). arXiv:2311.08581 [cs.CV]

Wojciech Zielonka, Timo Bolkart, and Justus Thies. 2023b. Instant volumetric head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4574–4584.