TEDRA: Text-based Editing of Dynamic and Photoreal Actors

Basavaraj Sunagad[1] Heming Zhu[1,†] Mohit Mendiratta[1,†]
Adam Kortylewski[1,3] Christian Theobalt[1,2] Marc Habermann[1,2,✉]
[1]Max Planck Institute for Informatics, Saarland Informatics Campus
[2]Saarbrücken Research Center for Visual Computing, Interaction and AI
[3]University of Freiburg

Abstract

Over the past years, significant progress has been made in creating photorealistic and drivable 3D avatars solely from videos of real humans. However, a core remaining challenge is the fine-grained and user-friendly editing of clothing styles by means of textual descriptions. To this end, we present TEDRA the first method allowing text-based edits of an avatar, which maintains the avatar's high fidelity, space-time coherency, as well as dynamics, and enables skeletal pose and view control. We begin by training a model to create a controllable and high-fidelity digital replica of the real actor. Next, we personalize a pre-trained generative diffusion model by fine-tuning it on various frames of the real character captured from different camera angles, ensuring the digital representation faithfully captures the dynamics of the real person. This two-stage process lays the foundation for our approach to dynamic human avatar editing. Utilizing this personalized diffusion model, we modify the dynamic avatar based on a provided text prompt using our Personalized Normal Aligned Score Distillation Sampling (PNA-SDS) within a model-based guidance framework. Additionally, we propose a time-step annealing strategy to ensure high-quality edits. Our results demonstrate a clear improvement over prior work in functionality and visual quality.

1. Introduction

Digital avatars of real humans play a vital role in various applications, including augmented and virtual reality, gaming, movie production, and synthetic data generation [10, 11, 13, 29, 30, 35, 45, 66, 72]. However, creating highly realistic and easily animatable avatars presents significant challenges due to the intricate and diverse nature of human geometry and appearance. While creating and editing highly quality avatars is possible, it remains a time-intensive and manual process requiring substantial expertise. Achieving a specific individual's likeness further adds complexity to this.

In the past years, text-driven image synthesis attracted the attention of the research community, as text is one of the most user-friendly data modalities that can be easily deployed without any expert knowledge. Thanks to the wide development and adaptation of transformers [47, 59] and diffusion models [49], several works have shown the ability to edit images in 2D [2, 50] and 3D [14, 23, 46], given a text prompt as input. While 2D-based methods produce visually convincing results, they most likely can not produce edits that are 3D-consistent. In contrast, 3D-based methods [14, 23, 46] show results on a 3D volume that can be rendered faithfully from an arbitrary camera viewpoint. However, such methods are mostly limited to static scenes.

Several methods have been proposed to generate 3D avatars from textual descriptions by distilling the 2D prior of generative models into 3D avatar representations [18, 21, 28]. However, despite the promising results, these methods often fail to adequately capture the dynamic and fine-grained details such as clothing movement. These aspects are crucial for interactive and dynamic applications.

Recent efforts [37, 52] have focused on modifying dynamic avatars while preserving their underlying motion. However, these approaches are restricted to the upper body [37, 52] and struggle to generalize to novel poses [37].

A significant and open challenge persists in seamlessly applying text-based edits to highly high-quality and controllable full-body avatars of real humans. The critical requirement is that these edits must maintain spatio-temporal consistency, dynamics, and the high fidelity of the original avatar, all while adhering to user-specified modifications.

In this work, we propose TEDRA, a novel text-based method to edit the appearance of a dynamic full-body avatar while preserving intricate details (see Fig. 1).

Our approach assumes a full-body avatar as input, which

† Equal contributions
✉ Corresponding author
Project page: vcai.mpi-inf.mpg.de/projects/Tedra

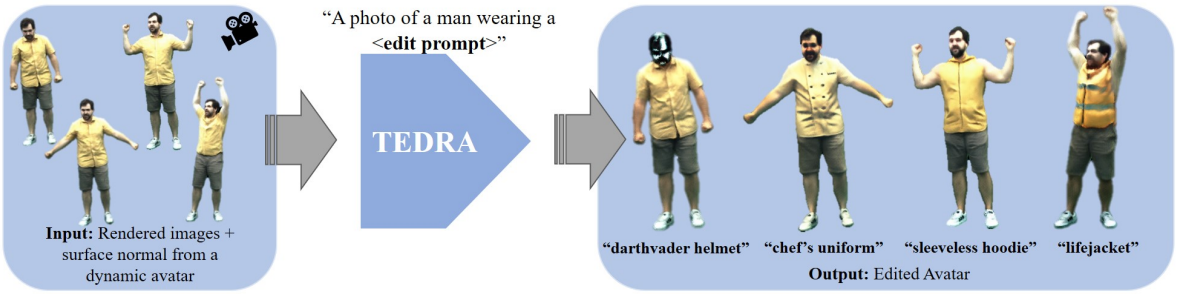

Figure 1. We propose a method for text-based editing of dynamic and photoreal actors (TEDRA). Our approach edits a pre-trained neural 3D human avatar according to a user-defined text prompt. Importantly, we preserve the original dynamics and view consistency of the digital avatar while also satisfying the desired edit.

is trained from multi-view video. In particular, our work builds upon TriHuman [74], where the avatar representation is modeled as a signed distance and radiance field anchored to an explicit and deformable mesh template, allowing fast inference of photorealistic human appearance and geometry. Through a pre-training stage, a drivable and photoreal digital avatar of the real actor is obtained.

Our method enables the text-based editing of such a dynamic volumetric avatar, ensuring spatial and temporal coherence. More precisely, we achieve editing through text-based conditional image generation utilizing a diffusion model. We contribute several technical advancements to ensure that the editing of TEDRA is authentic, personalized, and maintains visually convincing spatiotemporal consistency. Initially, we subsample frames from multi-view videos to capture the avatar's appearance and dynamics. We then fine-tune a pretrained diffusion model with these frames and a unique text identifier to create a personalized generative model that captures the avatar's detailed characteristics. Building on this, we introduce Personalized Normal-Aligned Score Distillation Sampling (PNA-SDS), a model-based, classifier-free method inspired by Zhang et al. [71]. The method employs two latent diffusion models—one personalized and one pre-trained—that perform iterative edits on the dynamic avatar. These diffusion models are conditioned on rendered normals of the avatar to preserve the dynamics while enhancing localized edits. Additionally, noise estimates from both the personalized and pre-trained diffusion models are strategically combined at specific timesteps to optimally balance the original avatar characteristics with the intended modifications.

Furthermore, to prevent oversaturation artifacts, we propose a windowed annealing strategy, that gradually reduces the influence of the personalized diffusion model enabling high-frequency edits. In summary, our contributions are:

- We introduce TEDRA, a method for editing dynamic full-body 3D avatars based on textual input. Our approach combines neural volumetric scene representations with text-driven diffusion models. This allows for precisely editing dynamic digital avatars while preserving detailed wrinkle patterns and ensuring seamless animatability.
- We propose Personalized Normal Aligned Score Distillation Sampling (PNA-SDS), a novel technique for personalized editing while preserving the integrity of avatar dynamics.
- We present windowed time-step annealing for score distillation from text-to-image diffusion models, which helps reduce oversaturation artifacts and supports detailed modifications.

We conduct a comprehensive evaluation of our method, employing both subjective and numerical assessments through a user study and comparisons with related techniques. The results demonstrate that our approach not only generates a wide range of text-based edits, but also maintains the integrity of the initial appearance. Additionally, we showcase animations of the edited avatars, further highlighting the temporal coherency and superior performance of our method compared to other relevant approaches.

2. Related Work

Diffusion-based Text-to-3D Generation and Editing. In the realm of computer graphics, the synthesis of 3D assets from textual descriptions has emerged as a captivating area of research. These assets are generated using Score Distillation Sampling (SDS), a technique introduced in DreamFusion [46], which enables the generation of 3D content from textual inputs by lifting text-to-image diffusion models to 3D domain. Fantasia3d [7] proposes a text-to-3D content creation method by disentangling the modeling and learning process of geometry and appearance. Meanwhile, Hi-FA [75] introduces a novel timestep annealing approach that progressively reduces the sampled timestep throughout a single-stage optimization process. InstructNerf2Nerf [14] employs an image-conditioned diffusion model for editing NeRF scenes with text instructions. Given a Neural Radiance Field (NeRF) [39] representation of a scene and respective multi-view images, the method uses an image-

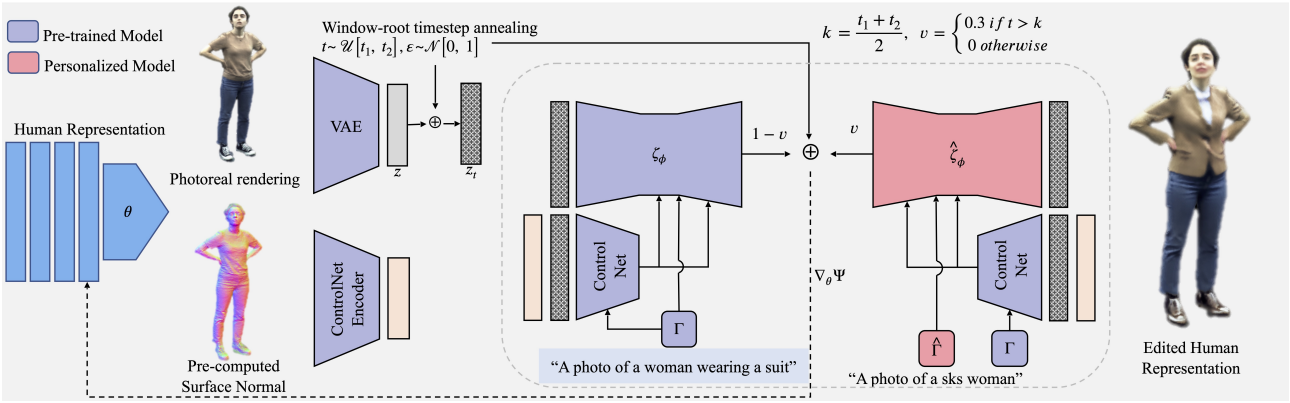

Figure 2. An overview of our approach for text-driven editing of dynamic and photoreal avatars (TEDRA). Our approach starts with a pre-trained TriHuman model as the base human representation. Then, we leverage a fine-tuned diffusion model in conjunction with our proposed Personalized Normal Aligned Score Distillation Sampling (PNA-SDS). The PNA-SDS technique then computes a normal aligned model-based score distillation sampling loss to optimize the human representation towards the edit prompt while preserving the subject's characteristics. This process is further enhanced by incorporating an annealing mechanism, which gradually refines the editing process.

conditioned diffusion model [3] to iteratively edit the input images while optimizing the underlying scene. In contrast, Re-PaintNerf [73] starts with the semantic selection of the object to be modified, followed by guiding the NeRF model using a pre-trained diffusion model. All these works have in common that they create *static* scenes, which neither support direct animation nor modeling of piece-wise rigid articulated objects and respective dynamics, e.g. motion-induced clothing deformations.

Diffusion-based Text-to-3D Avatar Generation. In the field of text-driven 3D avatar generation, a series of methodologies have been proposed recently, each tackling a specific challenge. For instance, AvatarVerse [68] and AvatarCraft [24] focus on generating avatars based solely on textual input. In contrast, DreamAvatar [4] prioritizes creating avatars with controllable poses and body types. HumanNorm [20] takes a unique approach, enhancing the perceived realism of 3D avatars by focusing on how 2D information translates to 3D geometry. Several methods, including DreamWaltz [21], DreamHuman [28], and TADA [32], combine text-driven generation with pre-built body models to create animatable avatars. Additionally, ZeroAvatar [65] and TeCH [22] aim to improve the overall quality and detail of generated avatars, while HaveFun [67] tackles the problem of reconstructing avatars from just a few photos. However, all the above works rely on a parametric body model to create a 3D avatar. When dealing with diverse avatars that deviate significantly from the parametric model, employing the original skinning weights in such instances will result in animations perceived as unrealistic. Secondly, and most importantly, they do not model the dynamics of the surface and appearance, i.e. wrinkles and cast shadows.

Recent works such as DynVideo-E [33] and Control4D [52] focus on 4D editing, ensuring temporal and spatial consistency by utilizing an underlying radiance field representation. However, these methods do not prioritize preserving the details and fidelity of the underlying avatar, nor do they adequately capture deformations such as wrinkles. A closely related work, AvatarStudio [37] introduces View-Time SDS for maintaining consistency in editing the underlying facial avatar across multiple views. Nevertheless, this method encounters challenges in generalizing to facial motions that it has not encountered before and faces limitations in real-time rendering. Additionally, the use of view-specific prompts in AvatarStudio leads to an entanglement of camera and identity views, restricting its applicability to datasets where the person's movement is limited.

Drivable 3D Avatars. Modeling high-fidelity, dynamic 3D-clothed human avatars has been an emerging topic in recent years. Here, we focus on the works related to the modeling of drivable 3D avatars, which take *only skeletal poses* and *virtual camera views* as input at inference time as they are most closely related.

Previous research on drivable avatars can be divided into two streams: mesh-based and hybrid approaches. Mesh-based methods [5, 12, 53, 60] represent the shape and appearance of dynamic characters with drivable template meshes with static (dynamic) textures. However, the rendering quality is bounded by the resolution of the underlying mesh template.

To improve the quality of both the generated geometry and rendering, hybrid approaches articulate implicit fields [15, 54, 55] or radiance fields [38, 69] with the explicit shape proxies, i.e., 3D skeletons, parametric human body models [26, 36, 42, 44], or person-specific template meshes [13, 29, 34, 74]. A prevalent research trend [1, 6, 9, 19, 25, 31, 41, 56, 57, 62, 64] focuses on modeling dynamic humans by mapping the posed space to a pose-agnostic canon-

ical space. To better model the pose-dependent appearance of humans, recent studies [10, 13, 29, 35, 45, 66, 72] incorporate motion-aware residual deformations in the canonicalized space. Among them, Neural Actor [35] and HD-Humans [13] leverage the texture space of the human body mesh as local features to model dynamic human appearances. Nevertheless, both methods require approximately 5 seconds to render a single frame. TriHuman [74] achieves real-time rendering and geometry generation through a deformable tri-plane anchored on the motion-controllable template mesh. The rendering and geometry quality is on par with, or even better than, the previous offline methods and significantly excels the real-time methods. We, therefore, take it as our underlying drivable avatar representation. Nevertheless, all these approaches exhibit a limitation as they do not allow for text edits and solely animate the clothing presented in the video.

3. Method

We aim to edit a pre-trained 3D human avatar, learned from multi-view video data, using textual prompts that specify desired changes, such as "Man wearing a hoodie" (see Fig. 2). The main challenge is to maintain the avatar's overall characteristics, transfer the original clothing dynamics, and achieve the specified edits.

We leverage the recent state-of-the-art neural 3D avatar representation, TriHuman [74], for its high geometric and visual quality and its real-time performance. The challenge lies in preserving intricate dynamics, such as wrinkles and other details, from the pre-trained human avatar during the editing process. To achieve this, we propose a novel score distillation-based approach, which effectively leverages the edit capabilities of large text-guided latent diffusion models (LDM) [8, 49] to coherently modify the motion-aware geometry and appearance generated by TriHuman. Next, before we introduce TEDRA (Sec. 3.2), we first discuss the necessary foundations, i.e., the avatar representation, LDMs, and Score Distillation Sampling (Sec. 3.1).

3.1. Preliminaries

Avatar Representation. Among the existing methods, we choose TriHuman [74] as the human representation for its ability to generate high-quality, motion-aware, and coherent geometry and appearance of dynamic humans in real-time.

As a hybrid representation, TriHuman firstly adopts an explicit, skeleton-driven human template mesh to depict the human character's coarse geometry. To generate the template mesh, it learns the non-rigid deformation of the human template mesh in the canonical space in a graph-to-graph translation manner [12]. The non-rigid deformed canonical template meshes are then transformed to the posed space via Dual Quaternion Skinning [27].

While the template mesh captures the motion-aware dynamics of the human characters, the fidelity of the appearance and geometry is bounded by the resolution of the template mesh. To this end, TriHuman introduced a deformable volume, parameterized with a Tri-plane, anchored on the deformable character's texture space. Given a spatial sample $\mathbf{x}_j$ along marching rays in the posed space, TriHuman non-rigidly maps the sample $\mathbf{x}_j$ to the texture volume bridged by the pose-deformed template mesh through inverse skinning. The respective position of $\mathbf{x}_j$ in the texture volume is denoted as $\mathbf{u}_j$. The sampled Tri-plane's features, denoted as $\mathbf{F}_{j,f}$ at position $\mathbf{u}_j$ are further fed into two shallow MLPs, i.e., the shape MLP $\mathcal{H}_{\text{sdf}}$, and the color MLP $\mathcal{H}_{\text{col}}$, to produce the corresponding SDF $s_{j,f}$ and color value $\mathbf{c}_{j,f}$.

$$\mathcal{H}_{\text{sdf}}(\mathbf{F}_{j,f}, p(\mathbf{u}_j) = s_{j,f}, \mathbf{q}_{j,f} \tag{1}$$

$$\mathcal{H}_{\text{col}}(\mathbf{q}_{j,f}, \mathbf{n}_{j,f}, p(\mathbf{d})) = \mathbf{c}_{j,f}. \tag{2}$$

where $\mathbf{q}_{j,f}$ denotes the motion-aware local shape features, $\mathbf{n}_{j,f}$ is the surface normal, p indicates positional encoding [38], f denotes the frame index, and $\mathbf{d}$ is the ray direction. Lastly, unbiased volume rendering [61] is adopted to integrate the ray samples and generate the final rendering. We refer to the original work [74] and our supplemental document for more details.

Latent Diffusion Models. In Latent Diffusion Models (LDMs) [49], an encoder function $\mathcal{E}$ maps a high-dimensional data point $\mathbf{x}$, e.g., an image, into a lower-dimensional latent representation $\mathbf{z} = \mathcal{E}(\mathbf{x})$. This latent space representation is then subject to a diffusion process [17, 40], a Markov chain of T steps. Each step adds a small amount of Gaussian noise, gradually transforming the data into a noise distribution. Mathematically, this process can be described as:

$$\mathbf{z}_t = \sqrt{\alpha_t}\mathbf{z} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(0, I). \tag{3}$$

where the diffusion timestep t ranges from $1 \leq t \leq T$. The reverse diffusion process in LDMs aims to gradually denoise the latent representation to generate new data samples. The reverse process is modeled by a neural network $\boldsymbol{\epsilon}_\phi$, where ϕ represents the trainable parameters of the network. This network learns to predict the noise $\boldsymbol{\epsilon}$ added at each step, enabling the model to reverse the diffusion process. Diffusion models, akin to various other generative models, are inherently capable of capturing conditional distributions denoted as $p(\mathbf{x}|\mathbf{c})$, where $\mathbf{c}$ represents a conditioning variable. The learning process for the conditional latent diffusion models then minimizes

$$\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,1),t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\phi(\mathbf{z}_t, t, \mathbf{c})\|_2^2\right]. \tag{4}$$

Score Distillation Sampling (SDS). Score Distillation Sampling [46] optimizes an implicit 3D representation from

textual descriptions to generate a view-consistent scene using pre-trained 2D text-to-image diffusion models. The approach constructs the scene through a differentiable image parameterization, employing a differentiable generator $\mathcal{G}$ that produces 2D images $\mathbf{x}$ from 3D scene parameters $\boldsymbol{\theta}$. The method utilizes a combination of a pre-trained and personalized diffusion model to derive a score function $\Psi_{\boldsymbol{\phi}}(\mathbf{x}_t, \mathbf{y}, t)$. This score function estimates the noise $\boldsymbol{\epsilon}$, given the noisy image $\mathbf{x}_t$, text embedding $\mathbf{y}$, and noise level t. The score function is crucial for determining the gradient direction for the scene parameter updates $\boldsymbol{\theta}$. The gradient, essential for updating these parameters, is computed as

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{SDS}}(\phi, \mathbf{x}) = \mathbb{E}_{t,\epsilon}\left[\mu(t)(\Psi_{\boldsymbol{\phi}}(\mathbf{x}_t; \mathbf{y}, t) - \boldsymbol{\epsilon})\frac{\partial \mathbf{x}}{\partial \boldsymbol{\theta}}\right]. \quad (5)$$

Here, $\mu(t)$ is a weighting based on the diffusion timestep t.

3.2. Proposed Method

We start by employing the pre-trained TriHuman model to generate images displaying diverse views and poses. Subsequently, we fine-tune the pre-trained Stable Diffusion model [48] on these generated images, utilizing an identity-specific prompt (Sec. 3.2.1). To ensure the accurate editing of the pre-trained avatar, we introduce our novel PNA-SDS loss (Sec. 3.2.2). This loss is computed through model-based classifier-free guidance [16], guided by the normal-aligned ControlNet [70]. Finally, we introduce our window-root timestep annealing strategy (Sec. 3.2.3), specifically designed for diffusion-guided text-to-3d editing.

3.2.1 Fine-tuning the Latent Diffusion Model

In dynamic full-body editing, the main challenge is preserving the original characteristics of the body, such as appearance, cloth deformations, and motion dynamics, rather than completely altering its appearance. DreamBooth [50] addresses identity-specific image generation by fine-tuning the Latent Diffusion Model (LDM) with an identity-specific prompt and a few images (4-5) of the subject. While effective for generating novel images, it is not suitable for consistently editing articulated and animatable avatars. Unlike AvatarStudio's [37] view-time fine-tuning scheme for short-sequence facial edits with 8-10 images, editing animatable full-body avatars is more complex and requires the LDM to manage novel poses and extensive view-dependent appearances, necessitating long sequences with a large number of frames for training. Furthermore, the fine-tuning strategy struggles with large datasets to generate accurate view-time token samples.

To address this, we propose a more comprehensive fine-tuning strategy that encompasses all conceivable poses and viewpoints, aiming to accurately represent the full spectrum of human surface dynamics and appearances. We render multi-view and multi-pose images denoted as $[\mathbf{x}_i; i \in \{1, ..., n\}]$ from the pre-trained TriHuman model for fine-tuning the UNet and the text encoder of the LDM. The fine-tuning process is further refined by incorporating an identity-specific token alongside a class noun within the prompt. This results in a structured prompt of the form 'a photo of a sks man/woman', where 'sks' is the identity-specific token.

We then fine-tune the UNet, denoted as $\hat{\zeta}_\phi$ and text encoder Γ, so the LDM can regenerate an image $\mathbf{x}_i$ from an initial noise map $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$ and a conditioning vector $\mathbf{s} = \Gamma(\mathbf{P})$, derived using a text encoder, conditioned on a text prompt $\mathbf{P}$. The fine-tuning of ζ_θ and Γ is supervised with a squared error loss function and a class-specific prior preservation loss [50]. The squared error loss for denoising a variably-noised image or latent code, given by $\mathbf{z}_{t,i} = \alpha_t \mathcal{E}(\mathbf{x}_i) + \beta_t \boldsymbol{\epsilon}$, is expressed as

$$\mathbb{E}_{x_i, s_i, \zeta, t}\left[w_t \|\hat{\zeta}_\phi(\mathbf{z}_{t,i}, \mathbf{s}) - \mathcal{E}(\mathbf{x}_i)\|_2^2\right], \quad (6)$$

where α_t, β_t, and w_t control the noise schedule.

After fine-tuning the LDM, we use a pre-trained normal-aligned ControlNet [70] encoder conditioned with a null prompt to control the virtual camera view and skeletal pose in the generated images.

3.2.2 Personalized Normal-aligned Model-based Score Distillation (PNA-SDS)

To enable subject-consistent, view, and pose-dependent edits, we compute the edit score with model-based classifier-free guidance [16] for updating the pretrained TriHuman model. This is done by interpolating the noise estimates from the fine-tuned model $\hat{\zeta}_\phi$ and pre-trained stable diffusion model ζ_ϕ, respectively, at specific timesteps. Our 3D human representation can generate surface normal maps, which guide the diffusion models via our normal-aligned pre-trained ControlNet. We adopt a parameter k based on the diffusion timestep t to determine whether or not to interpolate between the scores, which effectively balances between edits and appearance.

Given a rendered image $\mathbf{x}$, we obtain its latent representation $\mathbf{z}$ using the VAE of the stable diffusion: $\mathbf{z} = \mathcal{E}(\mathbf{x})$. We then uniformly sample a diffusion timestep $t \sim \mathcal{U}(t_1, t_2)$, where t_1 and t_2 represent the upper and lower limits of the diffusion noise timestep t. These limits are introduced in the next paragraph, where we discuss the annealing process. We apply this sampled timestep to introduce noise to the input latent, resulting in a noised latent $\mathbf{z}_t$.

Let $\mathbf{c}$ represent the embedding of the editing text prompt (e.g., 'a photo of a man wearing a hoodie'), $\hat{\mathbf{c}}$ represent the text embedding of the prompt for the identity (e.g., 'a photo of a sks man'), and $\mathbf{n}$ (Eq. 2) represents the normal map rendered with TriHuman. Then, we can obtain the editing score as follows:

$$\begin{aligned}\mathbf{\Psi}(\mathbf{z}_t, t, \mathbf{c}, \hat{\mathbf{c}}, \mathbf{n}) &= w\big((1-v)\,\boldsymbol{\zeta}_\phi(\mathbf{z}_t, \mathbf{c}, \mathbf{n}) + v\,\hat{\boldsymbol{\zeta}}_\phi(\mathbf{z}_t, \hat{\mathbf{c}}, \mathbf{n})\big) \\ &\quad + (1-w)\,\boldsymbol{\zeta}_\phi(\mathbf{z}_t, \mathbf{n}), \\ v &= \begin{cases} 0.3 & \text{if } t > k, \\ 0 & \text{otherwise.} \end{cases}\end{aligned} \tag{7}$$

where w is the overall guidance weight and v stands for the model guidance weight. We use Eq. 7 for the noise estimate and k represents a threshold timestep for using a personalized diffusion model. It is important to note that incorporating surface normals as a condition for both text-guided and null prompts is essential to maintain spatial relationships and surface orientations within the estimated score. Our Personalized Normal Aligned-SDS, combined with a comprehensive fine-tuning strategy, excels in generalizing edits to novel views and poses not seen in the fine-tuning dataset. In Fig. 5, we show the effectiveness of Personalized Normal Aligned-SDS compared with other SDS variants.

3.2.3 Windowed Root Timestep Annealing

Similar to previous studies [7, 63, 75], our experimental results indicate that Score Distillation Sampling (SDS) encounters notable challenges related to over-saturation and loss of fine details, particularly when a large timestep t is randomly selected. In diffusion models, the higher timesteps correspond to the semantics of the image, while the lower timesteps correspond to finer details [43]. Thus, for an editing task, it is crucial to establish the edit semantics early on and then move to add finer details.

Several annealing strategies have been proposed [7, 63, 75], but none of them are suitable for an editing task. HiFA [75] proposed a square root annealing strategy for selecting the diffusion timestep t based on the iteration step, directly correlating the diffusion process' progression with the training iteration. However, random sampling of t is crucial for model-based classifier-free guidance as per Eq. 7. When t is deterministically chosen, as in HiFA, we face the following limitations:

- With a fixed threshold $k = 600$, model-based guidance ceases after $t > k$, leading to a complete loss of original appearance due to the lack of influence from the personalized model in later iterations (refer to PNA-SDS+HiFA annealing in Sec. 4.3).
- Constant model-based guidance ($k = 0$) restricts edit flexibility and results in blurry artifacts.

Thus, it is important to stochastically sample t to balance the original appearance and edit flexibility.

To address this, we introduce a *windowed square root annealing strategy* specifically designed to modulate the annealing of timesteps while allowing random sampling. This approach ensures a more controlled and gradual progression of timesteps during the training process.

	InsN2N [14]	AvatarStudio [37]	**Ours**
Q1	9.6	12.8	**77.6**
Q2	12	24	**64**
Q3	14.4	8	**77.6**
Q4	11.2	10.4	**78.4**

Table 1. **User Study.** Results from our study with 25 participants. Our approach outperforms AvatarStudio and InstructNeRF2NeRF (InsN2N) by a large margin.

Given the total number of iterations N and the current iteration τ, along with a specified window size w, our annealing values are:

$$t_1 = t_{\text{max}} - (t_{\text{max}} - t_{\text{min}})\sqrt{\frac{\tau}{N}},\ t_2 = t_1 - w,\ k = \frac{t_1 + t_2}{2}. \tag{8}$$

Here, t_{max} and t_{min} represent the maximum and minimum diffusion timesteps, respectively. The window $[t_1, t_2]$ is the range within which t is randomly sampled. The parameter k signifies the timestep for using the personalized diffusion model. This dynamic window adapts throughout the annealing process, facilitating a balance between establishing edit semantics at higher timesteps and adding fine details at lower timesteps, thereby mitigating issues of over-saturation.

Moreover, the weight v of the personalized diffusion model is annealed to enhance the faithfulness of the edits to the input prompt. This approach not only improves the semantic integrity during initial higher timesteps but also ensures the preservation of fine details as the process progresses to lower timesteps.

4. Experiments

Our evaluation focuses on three key aspects: 1) Ensuring alignment with the target text prompt while maintaining the subject's inherent characteristics; 2) The capability to produce 3D consistent edits for high-quality free-view rendering; 3) The temporal coherency of the generated edits, enabling dynamic replay and skeleton animations.

Dataset. We conduct experiments on one subject (wearing shorts and a shirt) of the DynaCap dataset [12], which is recorded in a calibrated multi-camera studio. In addition, we captured three more subjects in various clothing in a similar studio setup. For training TriHuman, we obtain skeletal motion using markerless motion capture [58], and foreground masks using background matting [51]. For more details, we refer to the supplemental document.

4.1. Qualitative Results

Fig. 3 presents the text-based edits upon a single identity with various poses generated by TEDRA. The results obtained from TEDRA shine in the following aspects: **(1)**

Figure 3. **Qualitative Results.** We present the text-based visual editing results and the underlying geometry. Our method generates compelling text-driven visual edits, ensuring 3D and temporal consistency while altering appearance and geometry. We recommend the readers to **zoom in** for better viewing of the details.

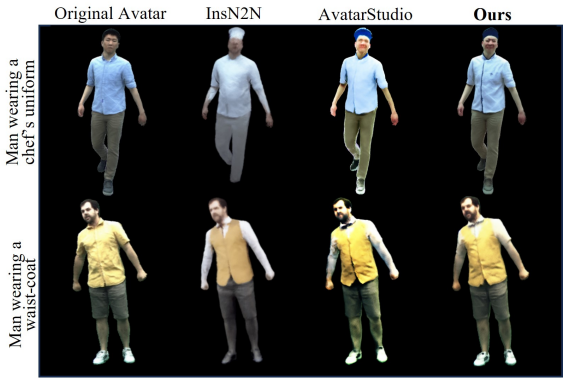

Figure 4. **Qualitative Comparison.** Our approach preserves the subject's characteristics and produces visually compelling edits that align with the prompts.

Localized and precise: As illustrated in Fig. 3 ("Women wearing sunglasses"), TEDRA yields fine-grained edits in the target region while preserving the clothing wrinkle details from the original avatar. **(2) Versatile and expressive:** As shown in Fig. 3 ("Women wearing a suit"), TEDRA can modify not only the appearance but also the style of the outfits. The edited outfits significantly differ from the originals in both, style and appearance. **(3) 3D and temporal coherent:** TEDRA edits, both the appearance and the underlying geometry of the clothed human avatar to align with the given text prompt, as evident in the representations of Fig. 3 ("Women wearing military uniform"). The edits seamlessly integrate with the drivable avatar and remain coherent across various poses.

4.2. Quantitative Evaluation

We compare our method with two recent 3D text-based editing approaches: InstructNeRF2NeRF [14] and AvatarStudio [37]. To ensure a fair comparison, we use the same training data for InstructNeRF2NeRF and apply AvatarStudio's full-head editing strategy to our full-body avatar. Fig. 4 shows the outcomes under different text prompts.

InstructNeRF2NeRF exhibits poor alignment with target prompts and significantly alters the original appearance by editing incorrect regions. AvatarStudio produces blurry results that lack fine details like wrinkles and clothing dynamics. In contrast, our method retains fine details and dynamics, producing more coherent edits aligned with the text.

Given the difficulty of quantitative comparison with ground truth in this setting, we adopted a user study approach from AvatarStudio [37]. Each session featured four side-by-side videos: the original input and randomly shuffled outputs from three text-driven methods. We evaluated three identities with three different prompts, totaling nine videos, each 10-15 seconds long. Participants were asked to respond to the following questions:

- Q1: Which method better preserves the appearance of the input sequence (subject consistency)?

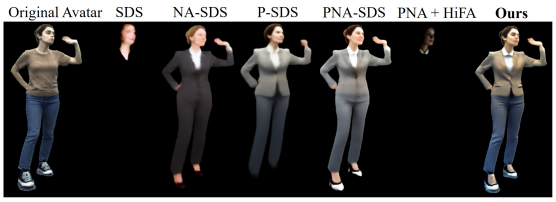

Figure 5. **Ablation Study.** Our full method achieves visually plausible edits and preserves the dynamics of the original character.

	InsN2N [14]	AvatarStudio [37]	**Ours**
CLIP Similarity ↑	0.1656	0.1412	**0.2098**
FID Score ↓	98.932	116.732	**80.74**

Table 2. We adopt CLIP text-image Direction Similarity to assess the alignment of the edits with the text within the CLIP space. Additionally, we report the FID scores to evaluate the preservation of appearance and geometric accuracy of the avatars.

- Q2: Which method better adheres to the provided textual prompt (prompt preservation)?
- Q3: Which method better maintains the animations and dynamics of the original motion (temporal consistency)?
- Q4: Which method performs better overall, considering the three aspects above?

Tab. 1 presents the user study from 25 participants, with our method consistently receiving the highest ratings. For overall quality (Q4), our method was preferred 78.4% of the time, compared to AvatarStudio's 10.4% and Instruct-NeRF2NeRF's 11.2%. We also evaluate CLIP Text-Image Direction Similarity and FID scores. Tab. 2 shows that our method outperforms prior works by a large margin.

4.3. Ablations

In Fig. 5 and Tab. 3, we conduct ablative studies to assess the effectiveness of our design choices. Initially, we establish the necessity of incorporating normals as a condition for, both, the edit-prompt and the null-prompt, we use a pre-trained diffusion model [48] and ControlNet [70] (Eq. 7 with $v = 0$), which we term as **NA-SDS**. As shown in Fig. 5, aligning the score with normals crucially helps to preserve geometric details that are lost with standard **SDS**.

Next, we highlight the importance of adopting the personalized model to preserve the appearance. Here, we compute the score using our proposed loss, but we exclude the normals to highlight the impact of the personalized model (Eq. 7 with $n = \oslash$) termed as **P-SDS**. We observe an improvement in the appearance and overall structure of the edits compared to vanilla **SDS**.

Although the model with Personalized-SDS, i.e, **P-SDS**, demonstrates significant improvements compared to standard **SDS**, the absence of geometric guidance from normals results in broken avatars. To this end, we use normal guidance in our **PNA-SDS** formulation Eq. 7, which further helps preserve the geometry as well as the key features in the original avatar. Yet, the random sampling of the timestep t leads to a diminished resolution of fine details, underscoring the necessity for an annealing process.

SDS	NA-SDS	P-SDS	PNA-SDS	PNA+HiFA	**Ours**
0.1153	0.2256	0.1844	0.2005	0.1159	**0.2409**

Table 3. CLIP-Similarity scores corresponding to the ablations presented in Fig.5. Note that each of our design choices leads to an improvement compared to the baselines.

Additionally, we illustrate the ineffectiveness of the annealing strategy proposed by HiFA [75] (termed as **PNA-SDS + HiFA annealing**) when applied to our method. Here, we set a constant value for k and deterministically select the value of diffusion timestep t based on the iteration step. This is ineffective for our approach as once $t < k$, the personalized model does not contribute to the score, resulting in a complete breakdown of the avatar.

In striking contrast, our full method, termed **Ours**, delivers high-quality edits while preserving the crucial details of the pre-trained human avatar. The CLIP scores for the ablation study (Tab. 3), further prove that our complete method outperforms all design alternatives in terms of alignment with the textual prompts and overall visual quality.

5. Conclusion

In this work, we tackled the problem of intuitive editing of 3D neural avatars where a user can specify a desired edit via text prompts. Our method then automatically adjusts the shape and appearance of the neural avatar to fulfill the user's demands while maintaining the subject's visual coherence. At the technical core, we propose a Personalized Normal-Aligned Score Distillation Sampling and a windowed timestep annealing to ensure space-time consistency in edits and high visual fidelity. Our results demonstrate a clear step towards more intuitive, high-fidelity 3D neural avatar editing and outperform respective competing methods. While TEDRA facilitates the intuitive editing of 3D neural human avatars with space-time consistency and superior visual quality, it faces challenges with long training times. Future work will focus on enabling more detailed edits while reducing resource demands.

Acknowledgements

Christian Theobalt was supported by ERC Consolidator Grant 4DReply (No.770784). Adam Kortylewski was supported by the German Science Foundation (No.468670075). This project was also supported by the Saarbrucken Research Center for Visual Computing, Interaction, and AI.

References

[1] Alexander Bergman, Petr Kellnhofer, Wang Yifan, Eric Chan, David Lindell, and Gordon Wetzstein. Generative neural articulated radiance fields. *NeurIPS*, 35:19900–19916, 2022. 3

[2] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 1

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 3

[4] 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

[5] Dan Casas, Marco Volino, John Collomosse, and Adrian Hilton. 4d video textures for interactive character appearance. *Comput. Graph. Forum*, 33(2):371–380, 2014. 3

[6] Jianchuan Chen, Ying Zhang, Di Kang, Xuefei Zhe, Linchao Bao, Xu Jia, and Huchuan Lu. Animatable neural radiance fields from monocular rgb videos. *arXiv preprint arXiv:2106.13629*, 2021. 3

[7] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 6

[8] CompVis. Stable diffusion, 2022. 4

[9] Yao Feng, Jinlong Yang, Marc Pollefeys, Michael J. Black, and Timo Bolkart. Capturing and animation of body and clothing from monocular video. In *SIGGRAPH Asia 2022 Conference Papers*, 2022. 3

[10] Qingzhe Gao, Yiming Wang, Libin Liu, Lingjie Liu, Christian Theobalt, and Baoquan Chen. Neural novel actor: Learning a generalized animatable neural representation for human actors. *IEEE TVCG*, 2023. 1, 4

[11] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5052–5063, 2020. 1

[12] Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Real-time deep dynamic characters. *ACM TOG*, 40(4), 2021. 3, 4, 6

[13] Marc Habermann, Lingjie Liu, Weipeng Xu, Gerard Pons-Moll, Michael Zollhoefer, and Christian Theobalt. Hdhumans: A hybrid approach for high-fidelity digital humans. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 6(3):1–23, 2023. 1, 3, 4

[14] Ayaan Haque, Matthew Tancik, Alexei Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 1, 2, 6, 7, 8

[15] Philipp Henzler, Niloy J Mitra, and Tobias Ritschel. Escaping plato's cave: 3d shape from adversarial rendering. In *CVPR*, pages 9984–9993, 2019. 3

[16] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. 5

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 4

[18] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *ACM Transactions on Graphics (TOG)*, 41(4):1–19, 2022. 1

[19] Shoukang Hu, Fangzhou Hong, Liang Pan, Haiyi Mei, Lei Yang, and Ziwei Liu. Sherf: Generalizable human nerf from a single image. *arXiv preprint*, 2023. 3

[20] 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*, 2023. 3

[21] Yukun Huang, Jianan Wang, Ailing Zeng, He Cao, Xianbiao Qi, Yukai Shi, Zheng-Jun Zha, and Lei Zhang. Dreamwaltz: Make a scene with complex 3d animatable avatars. 2023. 1, 3

[22] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. Tech: Text-guided reconstruction of lifelike clothed humans, 2023. 3

[23] Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. *CVPR*, 2022. 1

[24] Ruixiang Jiang, Can Wang, Jingbo Zhang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Avatarcraft: Transforming text into neural human avatars with parameterized shape and pose control, 2023. 3

[25] T. Jiang, X. Chen, J. Song, and O. Hilliges. In *CVPR*, pages 16922–16932, 2023. 3

[26] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *CVPR*, pages 8320–8329, 2018. 3

[27] Ladislav Kavan, Steven Collins, Jiří Žára, and Carol O'Sullivan. Skinning with dual quaternions. In *Proceedings of the 2007 symposium on Interactive 3D graphics and games*, pages 39–46, 2007. 4

[28] Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Fieraru, and Cristian Sminchisescu. Dreamhuman: Animatable 3d avatars from text. 2023. 1, 3

[29] Youngjoong Kwon, Lingjie Liu, Henry Fuchs, Marc Habermann, and Christian Theobalt. Deliffas: Deformable light fields for fast avatar synthesis. *NeurIPS*, 2023. 1, 3, 4

[30] Ruilong Li, Kyle Olszewski, Yuliang Xiu, Shunsuke Saito, Zeng Huang, and Hao Li. Volumetric human teleportation. In *ACM SIGGRAPH 2020 Real-Time Live!*, pages 1–1. 2020. 1

[31] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhofer, Jurgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. 2022. 3

[32] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxaing Tang, Yangyi Huang, Justus Thies, and Michael J Black.

Tada! text to animatable digital avatars. *arXiv preprint arXiv:2308.10899*, 2023. 3

[33] Jia-Wei Liu, Yan-Pei Cao, Jay Zhangjie Wu, Weijia Mao, Yuchao Gu, Rui Zhao, Jussi Keppo, Ying Shan, and Mike Zheng Shou. Dynvideo-e: Harnessing dynamic nerf for large-scale motion- and view-change human-centric video editing, 2023. 3

[34] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *NeurIPS*, 33: 15651–15663, 2020. 3

[35] 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 Trans. Graph.(ACM SIGGRAPH Asia)*, 2021. 1, 4

[36] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 3

[37] Mohit Mendiratta, Xingang Pan, Mohamed Elgharib, Kartik Teotia, Mallikarjun B R, Ayush Tewari, Vladislav Golyanik, Adam Kortylewski, and Christian Theobalt. Avatarstudio: Text-driven editing of 3d dynamic human head avatars. *ACM Trans. Graph.*, 42(6), 2023. 1, 3, 5, 6, 7, 8

[38] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 3, 4

[39] 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

[40] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models, 2021. 4

[41] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *ICCV*, 2021. 3

[42] Ahmed A. A. Osman, Timo Bolkart, and Michael J. Black. Star: Sparse trained articulated human body regressor. In *ECCV*, pages 598–613, 2020. 3

[43] Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. Localizing object-level shape variations with text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23051–23061, 2023. 6

[44] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, pages 10975–10985, 2019. 3

[45] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *ICCV*, pages 14314–14323, 2021. 1, 4

[46] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022. 1, 2, 4

[47] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018. 1

[48] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021. 5, 8

[49] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 4

[50] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. 2022. 1, 5

[51] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steve Seitz, and Ira Kemelmacher-Shlizerman. Background matting: The world is your green screen. In *CVPR*, 2020. 6

[52] Ruizhi Shao, Jingxiang Sun, Cheng Peng, Zerong Zheng, Boyao Zhou, Hongwen Zhang, and Yebin Liu. Control4d: Dynamic portrait editing by learning 4d gan from 2d diffusion-based editor. *arXiv preprint arXiv:2305.20082*, 2023. 1, 3

[53] Aliaksandra Shysheya, Egor Zakharov, Kara-Ali Aliev, Renat Bashirov, Egor Burkov, Karim Iskakov, Aleksei Ivakhnenko, Yury Malkov, Igor Pasechnik, Dmitry Ulyanov, et al. Textured neural avatars. In *CVPR*, pages 2387–2397, 2019. 3

[54] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhöfer. Deepvoxels: Learning persistent 3d feature embeddings. In *CVPR*, 2019. 3

[55] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *NeurIPS*, 2019. 3

[56] Shih-Yang Su, Frank Yu, Michael Zollhofer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. *NeurIPS*, 34: 12278–12291, 2021. 3

[57] Shih-Yang Su, Timur Bagautdinov, and Helge Rhodin. Danbo: Disentangled articulated neural body representations via graph neural networks. In *ECCV*, 2022. 3

[58] TheCaptury. The Captury. http://www.thecaptury.com/, 2020. 6

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 1

[60] Marco Volino, Dan Casas, John Collomosse, and Adrian Hilton. Optimal representation of multiple view video. In *BMVC*. BMVA Press, 2014. 3

[61] 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. 4

[62] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. Arah: Animatable volume rendering of articulated human sdfs. In *ECCV*, 2022. 3

[63] 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 6

[64] Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. Vid2actor: Free-viewpoint animatable person synthesis from video in the wild. *arXiv preprint arXiv:2012.12884*, 2020. 3

[65] Zhenzhen Weng, Zeyu Wang, and Serena Yeung. Zeroavatar: Zero-shot 3d avatar generation from a single image, 2023. 3

[66] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. *NeurIPS*, 34:14955–14966, 2021. 1, 4

[67] Xihe Yang, Xingyu Chen, Daiheng Gao, Xiaoguang Han, and Baoyuan Wang. Have-fun: Human avatar reconstruction from few-shot unconstrained images. *arXiv:2311.15672*, 2023. 3

[68] Huichao Zhang, Bowen Chen, Hao Yang, Liao Qu, Xu Wang, Li Chen, Chao Long, Feida Zhu, Kang Du, and Min Zheng. Avatarverse: High-quality & stable 3d avatar creation from text and pose, 2023. 3

[69] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 3

[70] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 5, 8

[71] Zhixing Zhang, Ligong Han, Arna Ghosh, Dimitris N. Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6027–6037, 2022. 2

[72] Zerong Zheng, Xiaochen Zhao, Hongwen Zhang, Boning Liu, and Yebin Liu. Avatarrex: Real-time expressive full-body avatars. *ACM TOG*, 42(4), 2023. 1, 4

[73] Xingchen Zhou, Ying He, F. Richard Yu, Jianqiang Li, and You Li. Repaint-nerf: Nerf editting via semantic masks and diffusion models, 2023. 3

[74] Heming Zhu, Fangneng Zhan, Christian Theobalt, and Marc Habermann. Trihuman: A real-time and controllable tri-plane representation for detailed human geometry and appearance synthesis. *ACM Trans. Graph.*, 2024. 2, 3, 4

[75] Junzhe Zhu and Peiye Zhuang. Hifa: High-fidelity text-to-3d generation with advanced diffusion guidance, 2023. 2, 6, 8