3DHumanEdit: Multi-modal Body Part-aware Conditioning Information Integration for 3D Human Manipulation

FeiFan Xu[1], Tianyi Chen[2*], Fan Yang[3], Yunfei Zhang[1], Si Wu[1]
[1]South China University of Technology
[2]City University of Hong Kong
[3]Nanyang Technological University
cs_feifan@mail.scut.edu.cn, tychen.cs@gmail.com, fan007@e.ntu.edu.sg
cszhangyunfei@mail.scut.edu.cn, cswusi@scut.edu.cn

Abstract

The rapid advancement of 3D Generative Adversarial Networks (GANs) has significantly enhanced the diversity and quality of generated 3D images. Despite these breakthroughs, the manipulation capabilities of 3D GANs remain unexplored, presenting substantial challenges for practical applications where user interaction and modification are essential. Current manipulation methods often lack the precision needed for fine-grained attribute manipulation, and struggle to maintain multi-view consistency during the editing process. To address these limitations, we propose 3DHumanEdit, a novel approach for 3D human body part-aware manipulation. 3DHumanEdit leverages multi-modal feature fusion and body part-aware feature alignment to achieve precise manipulation of individual body parts based on detailed text inputs and segmentation images. By exploring 3D prior for accurate editing and enforcing correspondence in latent space, 3DHumanEdit ensures coherence across multiple views. Experiments demonstrate that 3DHumanEdit outperforms existing methods in both editing fidelity and multi-view consistency, offering a robust solution for fine-grained 3D manipulation.

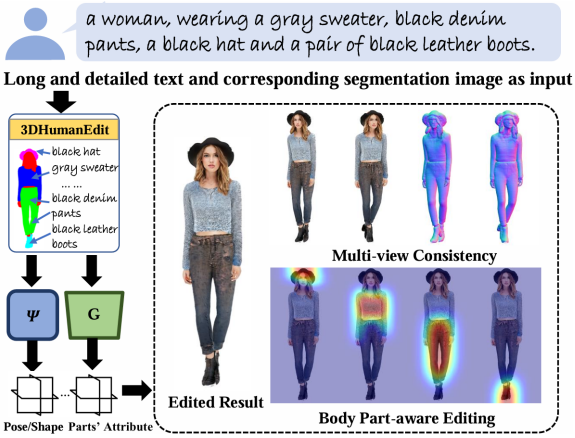

Figure 1: 3DHumanEdit generate 3D human bodies based on segmentation image and textual description in a body part-aware manner. The generated images exhibit high fidelity 3D consistency and detailed geometry.

Introduction

The increasing prevalence of 3D generative adversarial networks (GANs) in recent years has significantly advanced the 3D vision, enabling more realistic and varied representations. (Or-El et al. 2022; Chan et al. 2022, 2021). These developments have opened up new possibilities in various fields, including virtual reality, gaming, animation, and digital art (Oh and Jo 2024; Wu et al. 2021). However, research focusing on the editing capabilities of 3D GANs remains limited, posing significant challenges for practical applications where user interaction and modification are crucial. Existing 3D manipulation methods primarily rely on prompts to match and alter the appearance of 3D virtual avatars (Cao et al. 2024; Hong et al. 2022b). Although these methods have made some progress, they often fall short in terms of accuracy and flexibility. To better bind textual descriptions to specific attributes, approaches have been proposed that use example images to modify the latent space of NeRF (Mildenhall et al. 2021). These methods aim to enhance control over 3D models by establishing feature associations across different modalities. For instance, AvatarCLIP (Hong et al. 2022b) allows ordinary users to customize the shape and texture of 3D avatars and drive their actions based on textual descriptions using CLIP (Radford et al. 2021). However, this approach typically employs text features for coarse alignment of global characteristics, making it inadequate for capturing fine-grained details of individual body parts. Compositional Cross-modal Human (CCH) (Fu et al. 2023) utilizes cross-attention to link text descriptions with each part of the body and independently render each segment of the 3D human figure. Nevertheless, these attempts remain incomplete, lacking refined constraints for each body part and an exploration of manipulation consistency across multiple views. The main challenge in 3D human editing arises from the non-rigid geometric nature of the human body, which complicates the generation process (Yang et al. 2022). This complexity leads to inconsistencies in each body part across different views, making it difficult to reflect the user's intended modifications accurately. Ensuring that edits made to a 3D human object are consistently reflected in

*Corresponding author.

all other views is a complex task that requires sophisticated algorithms capable of handling intricate poses, shapes, and textures from multiple angles.

To address these challenges, we propose a novel approach named 3DHumanEdit, which is designed to overcome these limitations by leveraging advanced multi-modal feature fusion and body part-aware feature alignment. Multi-modal feature fusion enables cross-fusion between various modalities: textual and segmented images. This enhances the associations between text and fine-grained image regions and makes the more related manipulation result. Additionally, 3DHumanEdit applies body part-aware feature alignment, promoting a tighter connection between the editing text and each body part. Moreover, to ensure multi-view consistency, we augment the training set with both positive and negative samples and incorporate an additional estimator for 3D pose and shape to benefit the 3DHumanEdit training phase. 3DHumanEdit effectively maintains multi-view consistency while achieving fine-grained control over 3D human image manipulation. The main contributions of our proposed method are as follows:

- We facilitate fusion between multiple modalities by learning feature among texts and images, enabling fine-grained body part-aware manipulation of 3D images.
- 3DHumanEdit captures pose and shape with 3D human object, ensuring multi-view consistency in the manipulated images.
- 3DHumanEdit achieves optimal editing results compared to existing methods, enabling body part-aware fine-grained manipulation.

Related Work

3D-aware Human Generation

(Karras, Laine, and Aila 2019) In recent years, the rapid advancement of 3D generative models has transitioned from early techniques utilizing meshes (Liao et al. 2020; Gao et al. 2022), point clouds (Li et al. 2019), and voxel grids (Hao et al. 2021), to contemporary methods that represent generated 3D scenes using neural radiance fields and density fields. Researchers have combined implicit neural representations with GANs to learn 3D-aware image synthesis using only 2D images for supervision (Niemeyer and Geiger 2021; Or-El et al. 2022; Chan et al. 2022, 2021; Song et al. 2021, 2023). However, due to the non-rigid geometry of the human body, synthesizing 3D human images is more challenging (Yang et al. 2022). Consequently, exploration in this area is less extensive compared to generic generative models. Most transitional human generation works focus on 2D human generation conditioned by poses (Men et al. 2020; Yang and Lin 2021). Recently, StyleGAN-Human (Fu et al. 2022) expanded the human dataset and achieved high-quality 2D human generation based on StyleGAN2-ada(Karras et al. 2020).

In 3D-aware generation, most studies use predefined parametric human templates (Loper et al. 2015) to generate 3D human figures (Zhang et al. 2023a; Hong et al. 2022a; Noguchi et al. 2022; Bergman et al. 2022; Zhang et al. 2022, 2023b; Dong et al. 2023; Jiang et al. 2023). Notably, ENARF-GAN (Noguchi et al. 2022) proposed an unsupervised method to learn 3D human generation without direct pose supervision. GNARF (Bergman et al. 2022) introduced a surface-driven deformation method to handle distortions between different poses. However, these methods fail to produce high-quality human images. AvatarGen (Zhang et al. 2023a) improved generation quality through a deformation network modeling non-rigid dynamics. EVA3D (Hong et al. 2022a) proposed a composite human NeRF representation, dividing the human body into different local parts, rendering them separately, and combining them to achieve high-resolution human image generation. None of these methods support user-interactive editing of generated human figures. AG3D employs a multi-resolution cascaded discriminator, including a face detail discriminator, to generate high-quality human images. Furthermore, 3D-SGAN (Zhang et al. 2022) proposed a 3D-aware algorithm for human attribute generation, primarily extending the Giraffe framework (Niemeyer and Geiger 2021). 3D-SGAN models 3D human structure using coarse semantic masks in specific fixed poses and camera views, unable to generate whole views of human heads and only supports image generation with weak 3D constraints. SemanticHuman (Zheng et al. 2024) models humans and garments separately from a virtual try-on perspective, using a semantic-aware approach to establish triplane models for garments and rendering different resolution modules through a semantic renderer, thereby generating high-definition human images.

Text-Guided Human Manipulation

Generic 3D human generation models do not support interactive editing of generated humans or their attributes, typically using random noise or manually set latent codes for limited control over human images. From the onset of 2D generation, there have been various manipulation methods for interactive human image editing (Jiang et al. 2022; Huang et al. 2023; Morelli et al. 2023). A representative example is Text2Human (Jiang et al. 2022), which proposed using a codebook to encode text and appearance information, optimizing a VAE generator to enable corresponding manipulation by querying keys to extract values. In 3D-aware generation, methods such as (Hong et al. 2022b; Cao et al. 2023; Patashnik et al. 2021; Haque et al. 2023) have also explored using text-semantic embeddings to edit generated results. Cao et al. (Cao et al. 2023) employed a diffusion-based model with a cross-attention mechanism to process textual information, constraining the output 3D images with SDS loss for text-controlled avatar generation. While these methods have achieved good performance, they cannot achieve fully disentangled control of different fine-grained semantic parts and have long optimization times. CNeRF (Ma et al. 2023) used separate generation networks for each semantic category, achieving complete disentanglement of each semantic region, but the parallel training of multiple generators significantly increased computational costs, limiting its applicability in real-world scenarios. TADA (Liao et al. 2024) uses hierarchical rendering and SDS loss with a displacement layer and texture map to cre-

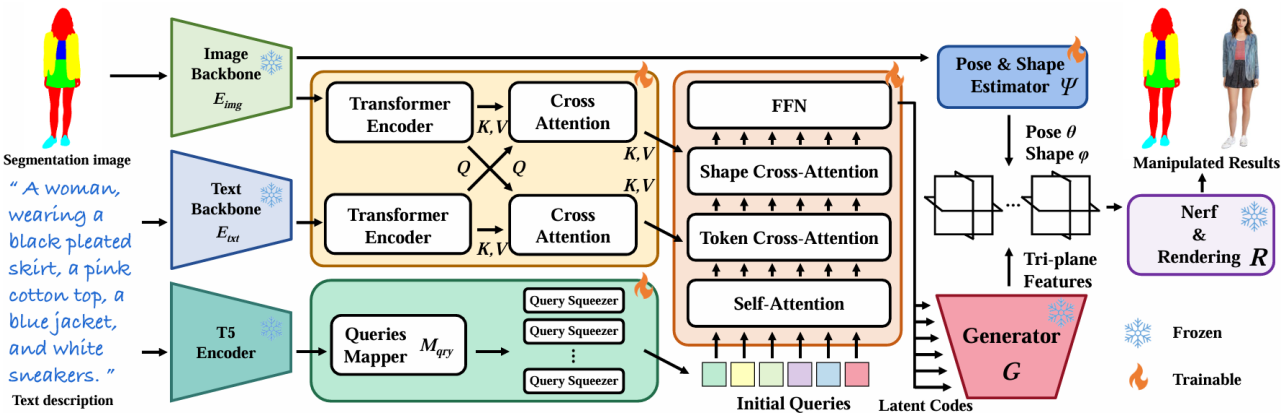

Figure 2: The flowchart of 3DHumanEdit. The proposed model receives two inputs: text description and segmentation image. The text descriptions are processed through a T5 encoder and aggregated using a mapping query M_{qry}. This aggregated representation interacts with the features obtained from the text via the text encoder E_{txt}, facilitating multi-modal feature fusion. The segmented image is processed through the image encoder E_{img} and interacts with the text to achieve body part-aware feature alignment. The additional estimator Ψ provides the corresponding shape and pose parameters to the 3D GAN. Together with the integrated latent representation as side input, these parameters are utilized in differentiable rendering to produce the final manipulated results.

ate realistic 3D human models. HumanNorm (Huang et al. 2024) introduces structural constraints through fine-tuned text-to-depth models. Recently, a cross-modal attention (Fu et al. 2023) was proposed to encode text information, align the features of the corresponding parts of the text and the image in the latent space, and perform fine-grained manipulation of the human body image. Our proposed model completely decouples each semantic category through text feature learning and the interaction between segmentations and tokens, thereby achieving high-quality 3D-Aware human images that can be arbitrarily manipulated.

Methodology

Data Pre-processing

3D human editing dataset consists of ground truth 2D human images x_{real} and corresponding segmentation images x_{segm}. We have consolidated the segmentation categories into six main parts: body, tops, outer, bottoms, shoes and accessories. For the corresponding text descriptions, we input the images x_{real} into a pretrained InstructBLIP model (Dai et al. 2023) to generate textual outputs, describing the features of the six parts, resulting in the corresponding text $t^{(1)}_{part}$ to $t^{(6)}_{part}$. We also construct a text template $\mathcal{T}$: A man/-woman '$t^{(1)}_{part}$', wearing '$t^{(2)}_{part}$', '$t^{(3)}_{part}$', '$t^{(4)}_{part}$', '$t^{(5)}_{part}$' and '$t^{(6)}_{part}$' to create complete sentences. Additionally, we label the dataset with images of the same identity from different viewpoints. Images with a yaw angle greater than 0 relative to the human center are considered positive samples x^{+}_{real}, and those with a yaw angle less than 0 are negative samples x^{-}_{real}. Thus, in the data preparation phase, we construct text descriptions for each category centered around the ground truth, text descriptions and multi-view training images.

Multi-modal Feature Fusion

Text descriptions serve as the input conditions for manipulating 3D human images. Intuitively, we choose the CLIP encoder to process both the entire description and segmented descriptions of individual body parts. However, research (Esser et al. 2024) indicates that CLIP ability to capture features of long texts is weaker compared to the T5 model (Raffel et al. 2020). Additionally, the sentences constructed by template $\mathcal{T}$ are excessively long relative to the descriptions of each body part. Therefore, we opted for the T5 encoder, which is an encoder-decoder model pre-trained on both unsupervised and supervised tasks, with each task converted into a text-to-text format. For long sentences $\mathcal{T}$, we can obtain T5 text embedding $f_{\mathcal{T}} = \text{T5}(\mathcal{T})$; similarly, for the segmented descriptions of the body parts, we can obtain the corresponding $f^{(i)}_{tokn} = E_t(t^{(i)}_{part})$ by CLIP text encoder E_{txt}.

Since the overall description ultimately needs to control various parts of the 3D human body, we perform tokenization of the T5 output features to finely encode the overall text semantics and align them with the body part feature semantics. We set up a 4-layer MLP queries mapper M_{qry} to obtain the corresponding semantic query encoding and several squeezers S_{qry} to compress the query data dimensionality:

$$f^{(i)}_{\mathcal{T}} = S^{(i)}_{qry}(M_{qry}(f_{\mathcal{T}})), \tag{1}$$

where (i) denotes the i-th body part. For the semantic encoding output $f^{(i)}_{\mathcal{T}}$, we first delve into the intrinsic relationships of T5 features, applying self-attention to the semantic encoding to get aggregated semantic queries. The attention module can be formally expressed as follows:

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) * V. \tag{2}$$

And the forward process can be formulated as follows:

$$\mathcal{W}_0^{(i)} = \text{Attn}(f_{\mathcal{T}}^{(i)}, f_{\mathcal{T}}^{(i)}, f_{\mathcal{T}}^{(i)}), \tag{3}$$

where the $\mathcal{W}_0^{(i)}$ denotes the initial queries. We input the segmentation images, which offer pixel-level positional information, into the CLIP image encoder E_{img} to obtain segmentation image features f_{segm}. To better integrate the prior knowledge of the image editing regions with the text, we introduce a cross-modal attention module. The segmentation image features serve as keys and values, while the segmented text features act as queries in cross-attention module. This setup enables the model to align and integrate segment textual descriptions with corresponding image features, ensuring that the image accurately reflects the segment words. We define the resulting fusion feature as *shape-correlated feature* f_s.

Conversely, the features $f_{tokn}^{(i)}$ are also used as keys and values, with the segmentation image features acting as queries. The cross attention mechanism ensures that textual features are enriched with corresponding visual context, further refining the text-semantic alignment. We then define the resulting fusion feature as *token-correlated feature* f_t. The forward equation (3) can be updated to:

$$\mathcal{W}_1^{(i)} = \text{Attn}(\mathcal{W}_0^{(i)}, f_t^{(i)}, f_t^{(i)}). \tag{4}$$

We further integrate the shape-correlated feature into the forward manipulation network, resulting in $\mathcal{W}_2^{(i)}$. The corresponding forward process is as follows:

$$\mathcal{W}_2^{(i)} = \text{Attn}(\mathcal{W}_1^{(i)}, f_s^{(i)}, f_s^{(i)}). \tag{5}$$

The shape-correlated information f_s and the token-correlated information f_t are fused respectively. After further processing, we obtain the final fusion, which is then aggregated using a feed-forward network to produce the final $\mathcal{W}_{final}^{(i)}$ as the 3D GAN side input for content manipulation. Then we define the symbols for the overall editing process. We denote the trainable editing network structure as Ω. Thus, the forward formula is as follows:

$$\mathcal{W}_{final}^{(i)} = \Omega(\mathcal{T}, t_{part}^{(i)}, x_{segm}). \tag{6}$$

And the overall optimization formula is as follows:

$$x_{edit} = \sum_i \mathcal{R}(G(\mathcal{W}_{final}^{(i)}), \theta, \varphi), \tag{7}$$

where $\mathcal{R}$ represents the differentiable rendering module, G denotes the 3D GAN's generator, θ and φ denote the pose and shape parameters of the SMPL model respectively.

Body Part-aware Feature Alignment

We previously explored feature fusion related to body parts, particularly the integration of segmentation images and text descriptions across different modalities. During the forward pass of the network, we design a multi-modal fusion mechanism to extract shape-correlated features f_s and token-correlated features f_t. This mechanism enables us to effectively combine information about body parts in the image with the corresponding textual descriptions. We aim to align the textual semantics with the body part that needs to be manipulated. For instance, when the description pertains to body of arm, the attention should focus on the arm area in the 3D image. To achieve this, we construct pixel-level mask by binarization of corresponding segmentation images, denoted as $\widetilde{x}_s$ and $\widetilde{x}_t$ respectively. x_s and x_t are trainable attention weight map of f_s and f_t attention module. The specific loss formulation is as follows:

$$\mathcal{L}^{attn} = \sum_i \left\| x_s^{(i)} - \widetilde{x}_s^{(i)} \right\|^2 + \left\| x_t^{(i)} - \widetilde{x}_t^{(i)} \right\|^2. \tag{8}$$

Further, features f_s and f_t are input into the manipulation network to obtain the corresponding latent representations $\mathcal{W}$. The manipulated results are derived from the color C output by the differentiable rendering process. The estimated colors C for rays r can be defined as follows:

$$C(r) = \int_{t_n}^{t_f} T(t)\sigma\left(r(t)\right) c(r(t), d) dt, \tag{9}$$

where $T(t)$ is the transmittance and $\sigma\left(r(t)\right)$ is the volume density at point $r(t)$.

3D-Aware Human Manipulation

Pose and Shape Estimator In 3D manipulation tasks, the goal is to control various 3D-related features, such as pose and shape, to facilitate the necessary adjustments in posture and changes in shape for target editing. The template's θ controls the pose parameters in the SMPL. Direct modification of these parameters is not user-friendly, as general users are unlikely to know the corresponding values. To address this, we designed a pose estimator Ψ that inputs the segmentation map corresponding to the textual input, identifies the pose features within it, and encodes them to obtain the corresponding $\theta = \Psi(x_{segm})$.

Additionally, considering that the body shape of the manipulated object may change, we ensure that the model to be aware of the subtle variations induced by shape. Therefore, we add a prediction branch to the pose estimator, making the new Pose and Shape Estimator (PSE) capable of predicting the body shape, resulting in the corresponding φ. The forward process of the PSE can be represented as:

$$\theta, \varphi = \Psi(x_{segm}). \tag{10}$$

By estimating the 3D-related parameters, 3DHumaneEdut can accurately derive the pose and shape parameters of 3D human body.

Multi-View Human Consistency During training, we ensure multi-view consistency by referencing edited images against positive samples (x_{real}^+) and negative samples (x_{real}^-). We imposed a regularization on the manipulation results, and ultilized the LPIPS metric and pixel-wise L2 loss to perform an constraint on the edited image x_{edit} and the ground truth x_{real} with each body part.

$$\mathcal{L}^{cons} = \sum_{\substack{x \in x_{real}, \\ x^+ real, x^- real}} \left[||\hbar(x_{edit}) - \hbar(x)||^2 + ||x_{edit} - x||^2 \right], \tag{11}$$

Figure 3: Multi-view manipulation results for 3DHumanEdit and competing methods.

where x_{edit} represent the edited output image corresponding to different views, $\hbar(\cdot)$ denotes the LPIPS feature output from VGG-16 (Simonyan and Zisserman 2014). By minimizing $\mathcal{L}^{cons}$, 3DHumanEdit learns to maintain multi-view consistency, making the edited human images more consistent across different viewpoints.

Loss and Optimization We also introduce CLIP loss to perform semantic constraints, which can be formulated as follows:

$$\begin{aligned}\mathcal{L}^{clip} &= E_{img}(x_{edit})^{\mathrm{T}} \otimes E_{img}(x) \\ &+ E_{img}(x_{edit})^{\mathrm{T}} \otimes E_{txt}(t),\end{aligned} \tag{12}$$

where the E_{img} and E_{txt} represent the image and text encoder in CLIP model, $\otimes$ denotes the matrix multiplication operator. To prevent degradation of the latent $\mathcal{W}$ in the GAN during training, we randomly sample latent $\mathcal{W}$ with 10^5 times to calculate the average $\overline{\mathcal{W}}$. We then compute the L2 loss for our final updated $\mathcal{W}^{(i)}_{final}$. The formulation $\mathcal{L}^{pena} = \sum_i ||(\overline{\mathcal{W}} - \mathcal{W}^{(i)}_{final})||^2$, serves as penalty term.

In summary, by integrating the aforementioned PSE Ψ and the 3D human manipulation network Ω, we can manipulate 3D poses and shapes by minimizing the objective function of Ψ and Ω:

$$\min_{\Psi,\Omega} \mathcal{L}^{attn} + \mathcal{L}^{cons} + \mathcal{L}^{clip} + \mathcal{L}^{pena}. \tag{13}$$

During network inference, by inputting the corresponding manipulation text and invoking the segmented image, the target manipulated image can be generated.

Experiments

Experimental Setup

Training Data and Metrics We generated 100,000 human body images using a 3D-aware generator (Zheng et al. 2024) trained on the DeepFashion dataset(Liu et al. 2016), where the generator can output RGB images and segmentation images. We then employed instructBLIP to generate descriptions based on the RGB images and corresponding prompts. These image and text data will serve as the training data for 3DHumanEdit. To assess the generated image quality, we employed the Frechet Inception Distance (FID) and Kernel Inception Distance (KID), which is a widely used metric in the image generation domain. In order to evaluate the similarity between the generated images and the corresponding text, we utilized the CLIP-Score, wherein various types of CLIP models were employed to obtain a more accurate estimation.

Comparison Methods We compare 3DHumanEdit with recent representative and state-of-the-art methods: Text2human(Jiang et al. 2022) employs a codebook to generate corresponding 2D human bodies based on text. DreamAvatar(Cao et al. 2023) utilizes the SMPL model and NeRF in conjunction with Score Distillation Sampling (SDS) (Poole et al. 2022) to produce 3D human bodies. Text2Avatar (Gong et al. 2024) utilizes a codebook-based multimodal encoding approach to transform textual descriptions into 3D human bodies. CCH (Fu et al. 2023) leverages EVA3D and attention to generate corresponding 3D human body. During the evaluation process, we relied on the official implementations or pretrained models sourced from the respective methods under consideration.

Qualitative Results

In Figure 3, we present the visualization results of 3DHumanEdit in comparison with competing methods. The experiments reveal several issues with the competing methods. Text2Human fails to maintain 3D consistency, resulting in semantic inconsistencies across different viewpoints. In some cases, images cannot be accurately manipulated

Methods	Quality		Relevance		
	FID↓	KID↓	CLIP-S(B)↑	CLIP-S(L)↑	CLIP-S(G)↑
Text2human	13.8122	0.0032	26.7142	22.0762	20.9753
DreamAvatar	21.3453	0.0045	29.5101	24.1323	24.7442
Text2Avatar	16.0214	0.0038	27.0461	23.5044	22.4453
CCH	15.0332	0.0036	27.8442	24.5152	23.7221
Ours	**11.4632**	**0.0029**	**31.3093**	**26.1144**	**25.1882**

Table 1: Quantitative comparison of 3DHumanEdit and competing methods in terms of image quality and semantic accuracy. The types of CLIP-Score (B), (L), and (G) correspond to CLIP models ViT-B/32, ViT-L/14, and ViT-L/14@336px respectively. Best results are **boldfaced** and second best results are underline.

by the text. Although DreamAvatar can maintain 3D consistency, it suffers from severe artifacts and exhibits the siamese phenomenon, where the character appears double-faced. Text2Avatar is able to roughly adhere to the semantic content of the textual description, yet its generated images exhibit significant distortion. CCH generally maintains 3D consistency and semantic alignment with the text, but there are slight errors in clothing colors. In terms of capturing fine-grained semantics such as the words 'Dark' and 'Tight', CCH is not able to control them well. In contrast, 3DHumanEdit generates more realistic images, achieving precise manipulation while maintaining 3D consistency. This demonstrates the superior performance of 3DHumanEdit in both image fidelity and semantic accuracy.

Quantitative Results

We conduct a quantitative comparison between 3DHumanEdit and competing methods. We randomly selected 10,000 textual descriptions and generated corresponding images using both 3DHumanEdit and the competing methods. The quality and accuracy of the generated results were evaluated using FID, KID, and CLIP-Score metrics, with all metrics computed under the same conditions to ensure a fair comparison. Table 1 presents the results of this quantitative comparison in terms of FID, KID, and CLIP-Score. First, we assess the fidelity of the generated images using FID and KID. DreamAvatar exhibit significant artifacts due to the SDS, resulting in a higher FID of 21.345. CCH achieves an FID of 15.0332, which can be attributed to its lack of 3D consistency constraints. 3DHumanEdit achieves the best FID of 11.463, attributed to the introduction of 3D consistency. In addition to image fidelity, semantic consistency is crucial for evaluating manipulation tasks. Text2Human and Text2Avatar, which utilize codebook-based approaches, struggle to understand complex textual content, resulting in CLIP-S(L) scores of 22.0762 and 23.504, respectively. 3DHumanEdit, benefiting from its body part-aware capabilities and efficient modality fusion method, demonstrates better control over human body features and superior semantic relevance, achieving a CLIP-S(L) score of 26.1144.

Body Part-aware Feature Alignment

In Figure 4, we conduct an experiment to demonstrate how textual features and image features association with each

Figure 4: Visualization of body part-aware feature alignment in attention mechanism.

Figure 5: Visual comparisons of different ablative models.

other, showcasing capability of body part-aware module . We extracted the corresponding attention weights from the 3DHumanEdit and visualized them on the images, revealing that the attention corresponding to specific words in the text focused accurately on the relevant regions of the images. For instance, when referring to a 'black hat', the model highlights the area of the image where the hat is located, indicating a strong correspondence between the textual description and visual representation. This attention mapping validates the model understanding of body part-aware features. The associated information captured during this attention process will be utilized in the subsequent fusion learning phase, enhancing the model to achieve precise control over the generated content.

Ablation Study

In Figure 5, we perform a series of ablation studies to verify the efficacy of each component proposed in 3DHumanEdit. The left image is the segmentation image and the subsequent images illustrate the effects of removing specific components from 3DHumanEdit. By disabling the 'PSE', the generated output without the shape information, resulting in less accurate body proportions. Disabling the text-correlated feature 'token' lacks the tokenization process, leading to a diminished understanding of the textual descriptions such as wrong boots and pants. Without T5 encoder, that means proposed model without long text feature, which negatively impacts linguistic coherence in the output. The image 'w/o CA' excludes the cross-attention mechanism, resulting in worse alignment between textual and visual features. Finally, the

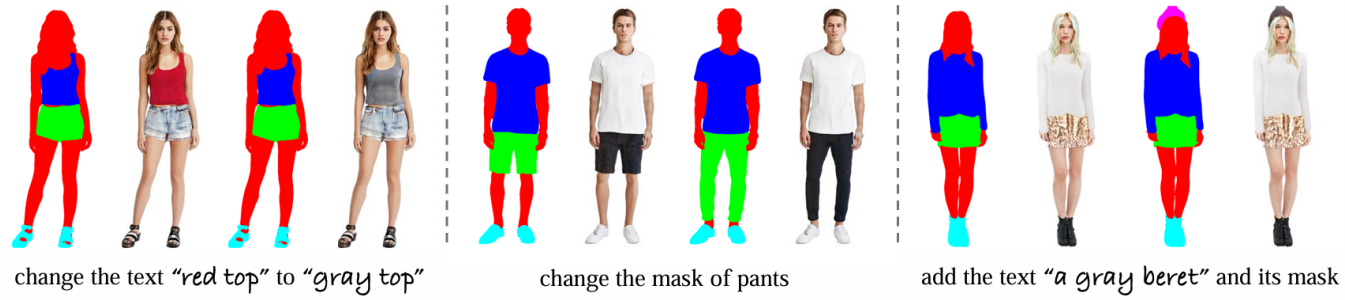

Figure 6: 3DHumanEdit enables local manipulation by optionally modifying segmentation image or text.

Methods	CLIP-S(B)↑	CLIP-S(L)↑	CLIP-S(G)↑
w/o Shape	26.1881	22.0422	22.8471
w/o Token	27.3537	23.6756	23.0038
w/o T5	28.8767	24.0788	23.8984
w/o CA	27.2666	22.4806	20.7435
Ours	**31.3093**	**26.1144**	**25.1882**

Table 2: Comparisons between 3DHumanEdit and its ablative models in terms of semantic accuracy.

Figure 7: Multi-view visual comparisons of CCH and 3DHumanEdit in the long text validation experiment.

image showcases the output from 3DHumanEdit, demonstrating the model capabilities in generating a coherent and contextually accurate representation of the description. This comparison highlights the importance of each component in achieving high-quality image generation. Further, Table 2 presents a comparison of 3DHumanEdit and its ablative models in terms of semantic accuracy. The model without shape information attained CLIP-S(B) scores, highlighting a significant drop in semantic coherence. Similarly, the model without text-correlated feature 'token' suggests that text-correlated feature plays a critical role in understanding the textual input. The omission of the T5 modelslightly improved the CLIP-related scores, but still fell short of the full model's performance. In contrast, 3DHumanEdit achieved the highest scores across all metrics, demonstrating its superior capability in generating semantically accurate representations. This highlights the importance of each component in maintaining high semantic correctness in the output.

Further Analysis

Figure 6 illustrates how 3DHumanEdit facilitates local manipulation by allowing users to modify segmentation images or text. For instance, users can change the text label of a garment from 'red top' to 'gray top', resulting in a corresponding modification of the top's color in the generated output. Similarly, users can directly adjust segmentation masks, such as altering the shape, color, or style of pants by manipulating the mask itself. Manipulated results capability to enhance creative control over image generation, enabling users to achieve precise and contextually relevant modifications. Figure 7 presents a multi-view visual comparison between CCH and 3DHumanEdit in the context of a long text validation experiment. The CCH captures the outfit but may lack certain nuances in detail and context. In contrast, 3DHumanEdit maintains the same outfit description but enhances the realism and coherence of the visual output. The variations highlight the model ability to create contextually rich representations, demonstrating its effectiveness in translating detailed textual descriptions into visually compelling images.

Conclusion

In this paper, we explore the manipulation capabilities of 3D GANs and address the challenges of achieving fine-grained and multi-view consistent attribute manipulation. We propose a novel approach that integrates multi-modal feature and body part-aware feature, enabling precise and 3D consistency manipulation of individual body parts based on detailed textual descriptions and segmentation images. Our experiments demonstrate the superiority of 3DHumanEdit over existing methods. This advancement presents a robust and effective solution for conducting fine-grained and body part-aware 3D Human manipulations.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Project No. 62072189), in part by the GuangDong Basic and Applied Basic Research Foundation (Project No. 2024A1515011437), and in part by TCL Science and Technology Innovation Fund (Project No. 20231752).

References

Bergman, A.; Kellnhofer, P.; Yifan, W.; Chan, E.; Lindell, D.; and Wetzstein, G. 2022. Generative neural articulated radiance fields. *Advances in Neural Information Processing Systems*, 35: 19900–19916.

Cao, Y.; Cao, Y.-P.; Han, K.; Shan, Y.; and Wong, K.-Y. K. 2023. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. *arXiv preprint arXiv:2304.00916*.

Cao, Y.; Cao, Y.-P.; Han, K.; Shan, Y.; and Wong, K.-Y. K. 2024. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 958–968.

Chan, E. R.; Lin, C. Z.; Chan, M. A.; Nagano, K.; Pan, B.; De Mello, S.; Gallo, O.; Guibas, L. J.; Tremblay, J.; Khamis, S.; 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.

Chan, E. R.; Monteiro, M.; Kellnhofer, P.; Wu, J.; and Wetzstein, G. 2021. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5799–5809.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv:2305.06500.

Dong, Z.; Chen, X.; Yang, J.; Black, M. J.; Hilliges, O.; and Geiger, A. 2023. AG3D: Learning to Generate 3D Avatars from 2D Image Collections. *arXiv preprint arXiv:2305.02312*.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*.

Fu, J.; Li, S.; Jiang, Y.; Lin, K.-Y.; Qian, C.; Loy, C. C.; Wu, W.; and Liu, Z. 2022. Stylegan-human: A data-centric odyssey of human generation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, 1–19. Springer.

Fu, T.-J.; Xiong, W.; Nie, Y.; Liu, J.; Oguz, B.; and Wang, W. Y. 2023. Text-guided 3D Human Generation from 2D Collections. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Gao, J.; Shen, T.; Wang, Z.; Chen, W.; Yin, K.; Li, D.; Litany, O.; Gojcic, Z.; and Fidler, S. 2022. Get3d: A generative model of high quality 3d textured shapes learned from images. *Advances In Neural Information Processing Systems*, 35: 31841–31854.

Gong, C.; Dai, Y.; Li, R.; Bao, A.; Li, J.; Yang, J.; Zhang, Y.; and Li, X. 2024. Text2Avatar: Text to 3d Human Avatar Generation with Codebook-Driven Body Controllable Attribute. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 16–20. IEEE.

Hao, Z.; Mallya, A.; Belongie, S.; and Liu, M.-Y. 2021. Gancraft: Unsupervised 3d neural rendering of minecraft worlds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14072–14082.

Haque, A.; Tancik, M.; Efros, A. A.; Holynski, A.; and Kanazawa, A. 2023. Instruct-nerf2nerf: Editing 3d scenes with instructions. *arXiv preprint arXiv:2303.12789*.

Hong, F.; Chen, Z.; Lan, Y.; Pan, L.; and Liu, Z. 2022a. Eva3d: Compositional 3d human generation from 2d image collections. *arXiv preprint arXiv:2210.04888*.

Hong, F.; Zhang, M.; Pan, L.; Cai, Z.; Yang, L.; and Liu, Z. 2022b. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535*.

Huang, X.; Shao, R.; Zhang, Q.; Zhang, H.; Feng, Y.; Liu, Y.; and Wang, Q. 2024. Humannorm: Learning normal diffusion model for high-quality and realistic 3d human generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4568–4577.

Huang, Z.; Chan, K. C.; Jiang, Y.; and Liu, Z. 2023. Collaborative diffusion for multi-modal face generation and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6080–6090.

Jiang, S.; Jiang, H.; Wang, Z.; Luo, H.; Chen, W.; and Xu, L. 2023. Humangen: Generating human radiance fields with explicit priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12543–12554.

Jiang, Y.; Yang, S.; Qiu, H.; Wu, W.; Loy, C. C.; and Liu, Z. 2022. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics (TOG)*, 41(4): 1–11.

Karras, T.; Aittala, M.; Hellsten, J.; Laine, S.; Lehtinen, J.; and Aila, T. 2020. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33: 12104–12114.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*.

Li, R.; Li, X.; Fu, C.-W.; Cohen-Or, D.; and Heng, P.-A. 2019. Pu-gan: a point cloud upsampling adversarial network. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7203–7212.

Liao, T.; Yi, H.; Xiu, Y.; Tang, J.; Huang, Y.; Thies, J.; and Black, M. J. 2024. Tada! text to animatable digital avatars. In *2024 International Conference on 3D Vision (3DV)*, 1508–1519. IEEE.

Liao, Y.; Schwarz, K.; Mescheder, L.; and Geiger, A. 2020. Towards unsupervised learning of generative models for 3d controllable image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5871–5880.

Liu, Z.; Luo, P.; Qiu, S.; Wang, X.; and Tang, X. 2016. Deepfashion: Powering robust clothes recognition and retrieval

with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1096–1104.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6): 1–16.

Ma, T.; Li, B.; He, Q.; Dong, J.; and Tan, T. 2023. Semantic 3D-aware Portrait Synthesis and Manipulation Based on Compositional Neural Radiance Field. *arXiv preprint arXiv:2302.01579*.

Men, Y.; Mao, Y.; Jiang, Y.; Ma, W.-Y.; and Lian, Z. 2020. Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5084–5093.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Morelli, D.; Baldrati, A.; Cartella, G.; Cornia, M.; Bertini, M.; and Cucchiara, R. 2023. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. In *Proceedings of the 31st ACM International Conference on Multimedia*, 8580–8589.

Niemeyer, M.; and Geiger, A. 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.

Noguchi, A.; Sun, X.; Lin, S.; and Harada, T. 2022. Unsupervised learning of efficient geometry-aware neural articulated representations. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, 597–614. Springer.

Oh, W.; and Jo, Y. 2024. From 2D Portraits to 3D Realities: Advancing GAN Inversion for Enhanced Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 737–746.

Or-El, R.; Luo, X.; Shan, M.; Shechtman, E.; Park, J. J.; and Kemelmacher-Shlizerman, I. 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.

Patashnik, O.; Wu, Z.; Shechtman, E.; Cohen-Or, D.; and Lischinski, D. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2085–2094.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140): 1–67.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Song, C.; Wei, J.; Li, R.; Liu, F.; and Lin, G. 2021. 3d pose transfer with correspondence learning and mesh refinement. *Advances in Neural Information Processing Systems*, 34: 3108–3120.

Song, C.; Wei, J.; Li, R.; Liu, F.; and Lin, G. 2023. Unsupervised 3d pose transfer with cross consistency and dual reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wu, X.; Zhang, Q.; Wu, Y.; Wang, H.; Li, S.; Sun, L.; and Li, X. 2021. F^3A-GAN: Facial Flow for Face Animation With Generative Adversarial Networks. *IEEE Transactions on Image Processing*, 30: 8658–8670.

Yang, F.; and Lin, G. 2021. Ct-net: Complementary transfering network for garment transfer with arbitrary geometric changes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9899–9908.

Yang, G.; Vo, M.; Neverova, N.; Ramanan, D.; Vedaldi, A.; and Joo, H. 2022. Banmo: Building animatable 3d neural models from many casual videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2863–2873.

Zhang, J.; Jiang, Z.; Yang, D.; Xu, H.; Shi, Y.; Song, G.; Xu, Z.; Wang, X.; and Feng, J. 2023a. Avatargen: a 3d generative model for animatable human avatars. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, 668–685. Springer.

Zhang, J.; Sangineto, E.; Tang, H.; Siarohin, A.; Zhong, Z.; Sebe, N.; and Wang, W. 2022. 3D-aware semantic-guided generative model for human synthesis. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, 339–356. Springer.

Zhang, X.; Zhang, J.; Chacko, R.; Xu, H.; Song, G.; Yang, Y.; and Feng, J. 2023b. Getavatar: Generative textured meshes for animatable human avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2273–2282.

Zheng, P.; Liu, T.; Yi, Z.; and Ma, R. 2024. SemanticHuman-HD: High-Resolution Semantic Disentangled 3D Human Generation. *arXiv preprint arXiv:2403.10166*.