Text-guided 3D Human Generation from 2D Collections

Tsu-Jui Fu¹, Wenhan Xiong², Yixin Nie²
Jingyu Liu², Barlas Oğuz², William Yang Wang¹

¹UC Santa Barbara ²Meta
{tsu-juifu, william}@cs.ucsb.edu {xwhan, ynie, jingyuli, barlaso}@meta.com

Abstract
3D human modeling has been widely used for engaging interaction in gaming, film, and animation. The customization of these characters is crucial for creativity and scalability, which highlights the importance of controllability. In this work, we introduce Text-guided 3D Human Generation (T³H), where a model is to generate a 3D human, guided by the fashion description. There are two goals: 1) the 3D human should render articulately, and 2) its outfit is controlled by the given text. To address this T³H task, we propose Compositional Cross-modal Human (CCH). CCH adopts cross-modal attention to fuse compositional human rendering with the extracted fashion semantics. Each human body part perceives relevant textual guidance as its visual patterns. We incorporate the human prior and semantic discrimination to enhance 3D geometry transformation and fine-grained consistency, enabling this to learn from 2D collections for data efficiency. We conduct evaluations on DeepFashion and SHHQ with diverse fashion attributes covering the shape, fabric, and color of upper and lower clothing. Extensive experiments demonstrate that CCH achieves superior results for T³H with high efficiency.

1 Introduction
Our world is inherently three-dimensional, where this nature highlights the importance of 3D applications in various fields, including architecture, product design, and scientific simulation. The capability of 3D content generation helps bridge the gap between physical and virtual domains, providing an engaging interaction within digital media. Furthermore, realistic 3D humans have vast practical value, especially in gaming, film, and animation. Despite enhancing the user experience, the customization of the character is crucial for creativity and scalability. Language is the most direct way of communication. If a system follows the description and establishes

---

Project website: https://text-3dh.github.io

![Image of 3D humans with fashion descriptions]

The man is sporting a short-sleeved cotton t-shirt and pure-colored medium shorts
The lady is wearing a graphic-patterned tank top, paired with three-point denim pants
The woman is wearing a floral-patterned long-sleeved sweater and long denim pants
She is dressed in a solid-colored sleeveless shirt with long cotton trousers

Figure 1: Text-guided 3D Human Generation (T³H).

the 3D human model, it will significantly improve controllability and meet the considerable demand.

We thus introduce Text-guided 3D Human Generation (T³H) to generate a 3D human with the customized outfit, guided via the fashion description. Previous works (Kolotouros et al., 2019; Gao et al., 2022) depend on multi-view videos to learn 3D human modeling, but these data are difficult to obtain and are not language controllable. Text-to-3D (Jain et al., 2022; Poole et al., 2023) has shown attractive 3D generation results through the success of neural rendering (Mildenhall et al., 2020). However, these methods apply iterative inference optimization by external guidance, which is inefficient for usage.

To tackle these above issues, we propose Compositional Cross-modal Human (CCH) to learn T³H from 2D collections. CCH divides the human body into different parts and employs individual volume rendering, inspired by EVA3D (Hong et al., 2023).
We extract the fashion semantics from the description and adopt cross-modal attention to fuse body volumes with textual features, where each part can learn to perceive its correlated fashion patterns. To support various angles of view, CCH leverages the human prior (Bogo et al., 2016) to guide the geometry transformation for concrete human architecture. Then these compositional volumes can jointly render a 3D human with the desired fashion efficiently. The semantic discrimination further considers compositional distinction over each human part, which improves the fine-grained alignment with its description through adversarial training.

We perform experiments on DeepFashion (Liu et al., 2016; Jiang et al., 2022) and SHHQ (Fu et al., 2022a), which contain human images with diverse fashion descriptions. The patterns include various types of shapes (sleeveless, medium short, long, etc.), fabrics (denim, cotton, furry, etc.), and colors (floral, graphic, pure color, etc.) for the upper and lower clothing. To study the performance of T3H, we conduct a thorough evaluation from both visual and semantic aspects. We treat overall realism, geometry measure, and pose correctness to assess the quality of generated 3D humans. For the alignment with the assigned fashion, we apply text-visual relevance from CLIP and fine-grained accuracy by a trained fashion classifier.

The experiments indicate that language is necessary to make 3D human generation controllable. Our proposed CCH adopts cross-modal attention to fuse compositional neural rendering with textual fashion as 3D humans, and semantic discrimination further helps fine-grained consistency. In summary, our contributions are three-fold:

- We introduce T3H to control 3D human generation via fashion description, learning from collections of 2D images.
- We propose CCH to extract fashion semantics in the text and fuse with 3D rendering in one shot, leading to an effective and efficient T3H.
- Extensive experiments show that CCH exhibits sharp 3D humans with clear textual-related fashion patterns. We advocate that T3H can become a new field of vision-and-language research.

2 Related Work

**Text-guided Visual Generation.** Using human-understandable language to guide visual generation can enhance controllability and benefit creative visual design. Previous works built upon adversarial training (Goodfellow et al., 2015; Reed et al., 2016) to produce images (Xu et al., 2018; El-Nouby et al., 2019; Fu et al., 2020, 2022c) or videos (Marwah et al., 2017; Li et al., 2018b; Fu et al., 2022b) conditioned on given descriptions. With sequential modeling from Transformer, vector quantization (Esser et al., 2021) can generate high-quality visual content as discrete tokens (Ramesh et al., 2021; Ding et al., 2021; Fu et al., 2023). The denoising diffusion framework (Ho et al., 2020; Ramesh et al., 2022; Saharia et al., 2022; Feng et al., 2023) gains much attention as its diversity and scalability via large-scale text-visual pre-training. Beyond images and videos, 3D content creation is more challenging due to the increasing complexity of the depth dimension and spatial consistency. In this paper, we consider text-guided 3D human generation (T3H), which has vast applications in animated characters and virtual assistants.

**3D Generation.** Different representations have been explored for 3D shapes, such as mesh (Gao et al., 2019; Rash et al., 2020; Henderson et al., 2020), voxel grid (Tatarchenko et al., 2017; Li et al., 2017), point cloud (Li et al., 2018a; Yang et al., 2019; Luo et al., 2021), and implicit field (Chen and Zhang, 2019; Park et al., 2019; Zheng et al., 2022). Neural Radiance Field (NeRF) (Mildenhall et al., 2020; Barron et al., 2022; Muller et al., 2022) has shown remarkable results in novel view synthesis (Schwarz et al., 2021; Chan et al., 2021; Skorokhodov et al., 2023) and 3D reconstruction (Yariv et al., 2021; Zhang et al., 2021). With the differentiable neural rendering, NeRF can be guided by various objectives. Text-to-3D draws appreciable attraction these days, which adopts external text-visual alignments (Jain et al., 2022; Khalid et al., 2022; Michel et al., 2022; Wang et al., 2022a; Hong et al., 2022) and pre-trained text-to-image (Wang et al., 2022b; Nam et al., 2022; Poole et al., 2023; Metzer et al., 2023; Tang et al., 2023; Seo et al., 2023). However, existing methods take numerous iterations to optimize a NeRF model, which is time-consuming for practical usage. Our CCH learns to extract fashion semantics with NeRF rendering and incorporates the human prior for a concrete human body, achieving effective and efficient T3H.

**3D Human Representation.** To reconstruct a 3D human, early works (Collet et al., 2015; Guo et al., 2019; Su et al., 2020) count on off-the-shelf tools to predict the camera depth. As mitigating the costly hardware requirement, they estimate a 3D human
We present text-guided 3D human generation (T3H) via the UV mapping (Shysheya et al., 2019; Yoon et al., 2021). With the promising success of NeRF, recent works (Peng et al., 2021b,a; Su et al., 2021) adopt volume rendering for 3D humans from multi-view videos (Weng et al., 2022; Chen et al., 2022). Since the data are difficult to collect, the 3D-aware generation (Chan et al., 2022; Gu et al., 2022; Noguchi et al., 2022) learns 3D modeling from the collection of human images (Yang et al., 2022; Hong et al., 2023). In place of arbitrary outputs, we introduce the first controllable 3D human generation that also learns from a 2D collection, and the presented fashion patterns should align with the description.

3 Text-guided 3D Human Generation

3.1 Task Definition

We present text-guided 3D human generation (T3H) to create 3D humans via fashion descriptions. For data efficiency, a 2D collection $D=\{V, T\}$ is provided, where $V$ is the human image, and $T$ is its fashion description. Our goal is to learn the neural rendering that maps $T$ into an articulate 3D human with the fashion patterns of $V$.

3.2 Background

Neural Radiance Field (NeRF) (Mildenhall et al., 2020) defines implicit 3D as $\{c, \sigma\} = F(x, d)$. The query point $x$ in the viewing direction $d$ holds the emitted radiance $c$ and the volume density $\sigma$. To get the RGB value $C(r)$ of certain rays $r(t)$, volume rendering is calculated along a ray $r$ from the near bound $t_n$ to the far bound $t_f$:

$$T(t) = \exp(- \int_{t_n}^{t} \sigma(r(s))ds),$$

$$C(r) = \int_{t_n}^{t_f} T(t) \sigma(r(t))c(r(t), d)dt, \quad (1)$$

where $T(t)$ stands for their accumulated transmittance. StyleSDF (Or-El et al., 2022) then replaces $\sigma$ with single distance field (SDF) $d(x)$ for a better surface, where $\sigma(x) = \alpha^{-1}\text{sigmoid}(\frac{-d(x)}{\alpha})$ and $\alpha$ is a learnable scalar that controls the tightness of the density around the surface boundary.

SMPL (Bogo et al., 2016) defines the human body as $\{\beta, \theta\}$, where $\beta \in \mathbb{R}^{10}$ and $\theta \in \mathbb{R}^{3 \times 23}$ control its shape and pose. We consider Linear Blend Skinning (LBS) as the transformation from the canonical into the observation space for the point $x$ to $\sum_{k=1}^{K} h_k H_k(\theta, J)x$, where $h_k \in \mathbb{R}$ is the scalar of the blend weight and $H_k \in \mathbb{R}^{4 \times 4}$ is the transformation matrix of the $k$th joint. Inverse LBS transforms the observation back to the canonical space as a similar equation but with an inverted $H$.

3.3 Compositional Cross-modal Human

Following EVA3D (Hong et al., 2023), we split the human body into 16 parts. As shown in Fig. 2, each body part holds its own bounding box $\{o_{\beta_{\text{min}}}', o_{\beta_{\text{max}}}'\}$. To leverage the human prior for a target pose $\theta$, we transform these pre-defined bounding boxes with SMPL’s transformation matrices $H_k$. Ray $r(t)$ is sampled for each pixel on the canvas. For a ray that intersects bounding boxes, we pick up its near and far bounds $(t_n$ and $t_f$) and sample $N$ points as follows: $t_i \sim U\left[\frac{1}{N} t_n + \frac{i}{N} t_f - t_n, \frac{N}{N} t_f - t_n\right]$.
We then transform these sampled points back to the canonical space with inverse LBS. For shape generalization, we consider not only pose transformation but also blend shapes \( B^P(\theta) \) and \( B^S(\beta) \) (Zeng et al., 2021). \( \mathbb{V} \) contains \( K \) nearest vertices \( v_k \) of the target SMPL mesh for the sample point ray \( r(t_i) \):

\[
\begin{align*}
g_k &= \frac{1}{||r(t_i) - v_k||^2}, \\
M_k &= \left( \sum_{k=1}^{K} g_k H_k \right) \begin{bmatrix} I B^P_k + B^S_k \end{bmatrix}, \\
x_i &= \sum_{v_k \in \mathbb{V}} \sum_{v_k \in \mathbb{V}} g_k (M_k)^{-1} \begin{bmatrix} r(t_i) \end{bmatrix}
\end{align*}
\]

where \( g_k \in \mathbb{R} \) is the inverse weight of the vertex \( v_k \) and \( M_k \in \mathbb{R}^{4 \times 4} \) is the transformation matrix. The \( x_i \) can be used for further volume rendering.

**Cross-modal Attention.** During rendering, if the canonical point \( x_i \) with the viewing direction \( d_i \) is inside the \( b \) th bounding box, it will be treated as:

\[
\begin{align*}
\hat{x}^b_i &= 2x_i - (o^b_{\text{max}} + o^b_{\text{min}}), \\
f^b_i &= \text{Linear}(\hat{x}^b_i, d_i), 
\end{align*}
\]

where a linear mapping is applied to acquire preliminary features \( f^b_i \in \mathbb{R}^{16 \times 8 \times 128} \). To exhibit the desired fashion in the final rendering, we extract the word features by the text encoder as \( \{ w_l \in \mathbb{R}^{512} \} \) from \( \mathcal{T} \). We then fuse the textual features with \( f^b_i \) via cross-modal attention:

\[
\begin{align*}
p_l &= \frac{\exp(f^b_i W^b w^T_l)}{\sum_{l=1}^{L} \exp(f^b_i W^b w^T_l)}, \\
CA(f^b_i | \{ w \}) &= \sum_{l=1}^{L} p_l w_l, 
\end{align*}
\]

where \( L \) is the length of \( \mathcal{T} \) and \( W^b \) is the learnable matrix. In this way, each point can learn to perceive relevant textual guidance for the \( b \) th human body part and depict corresponding fashion patterns.

Each body part has its individual volume rendering \( F^b \), which consists of stacked multilayer perceptrons (MLPs) with the SIREN activation (Sitzmann et al., 2020). Since the point \( x_i \) may fall into multiple boxes \( B_i \), we follow EVA3D to apply the mixture function (Lombardi et al., 2021):

\[
\begin{align*}
\{ c_i, \sigma_i \} &= F^b(\text{CA}(x^b_i, d_i | \{ w \})), \\
u_b &= \exp(-m(\hat{x}^b_i x_i^n + \hat{x}^b_i y_i^n + \hat{x}^b_i z_i^n)), \\
\{ c_i, \sigma_i \} &= \frac{1}{\sum_{b \in \mathbb{B}} u_b} \sum_{b \in \mathbb{B}} u_b \{ c_i, \sigma_i \}, 
\end{align*}
\]

where \( m \) and \( n \) are hyperparameters. With \( \{ c_i, \sigma_i \} \), we adopt Eq. 1 to render the RGB value of ray \( r(t_i) \). Through all sampled rays \( r \), we then have our final human rendering \( R \), where the overall process can be simplified as \( R = G(\beta, \theta | \{ T \}) \). To summarize, CCH leverages the human prior and adopts inverse LBS to acquire the canonical space for the target pose. The human body is divided into 16 parts, and each of them fuses its correlated fashion semantics via cross-modal attention. Finally, compositional bodies jointly render the target 3D human.

**Semantic Discrimination.** With the SMPL prior, our CCH contains robust geometry transformation for humans and can learn from 2D images without actual 3D guidance. For a ground-truth \( \{ \mathcal{V}, \mathcal{T} \} \), we parse the 2D human image as the segmentation \( \mathcal{S} \) (MMHuman3D, 2021), which provides the reliable body architecture. To obtain its fashion map \( Q \), we apply cross-modal attention between \( \mathcal{S} \) and \( \mathcal{T} \):

\[
\{ e_{i,j} \} = \text{Conv}(\mathcal{S}), \\
Q_{i,j} = \sum_{l=1}^{L} \frac{\exp(e_{i,j} W w^T_l)}{\sum_{l=1}^{L} \exp(e_{i,j} W w^T_l)} w_l, 
\]

where \( e \) is the same dimension as \( f, W \) is the learnable attention matrix, and \( Q \) perceives which human body part should showcase what fashion patterns. We concatenate the rendered human \( R \) (or the ground-truth \( \mathcal{V} \)) with \( Q \) and feed them into our discriminator \( D \) to perform binary classification:

\[
D(R | \{ T \}) = \text{BC}(\{ \text{Conv}(R), Q \}).
\]

In consequence, \( D \) can provide alignments of both

\[\text{Algorithm 1: Compositional Cross-modal Human}\]

1. \( D \): 2D collection of human images / fashion descriptions
2. \( G, D \): the generator / discriminator model for T3H
3. while TRAIN, CCH do
4. \( \mathcal{V}, \mathcal{T} \leftarrow \text{sampled human / description from} \mathcal{D} \)
5. \( \{ \beta, \theta \} \leftarrow \text{estimated SMPL parameters of} \mathcal{V} \)
6. \( \{ x_i \} \leftarrow \text{canonical points via inverse LBS} \quad \triangleright \text{Eq. 2, Eq. 4} \)
7. \( f_i \leftarrow \text{rendering features inside the} \ b \text{th box} \quad \triangleright \text{Eq. 3, Eq. 5} \)
8. \( \{ w_l \} \leftarrow \text{extracted textual features of} \mathcal{T} \quad \triangleright \text{Eq. 4} \)
9. \( \mathcal{CA} \leftarrow \text{fusion via cross-modal attention} \quad \triangleright \text{Eq. 4} \)
10. \( \{ c_i, \sigma_i \} \leftarrow \text{mixture radius / density} \quad \triangleright \text{Eq. 5} \)
11. \( R \leftarrow \text{final rendering human} \quad \triangleright \text{Eq. 4} \)
12. \( \mathcal{S} \leftarrow \text{segmentation map of} \mathcal{V} \quad \triangleright \text{Eq. 6} \)
13. \( Q \leftarrow \text{fashion map between} \mathcal{S} \text{and} \ \mathcal{T} \quad \triangleright \text{Eq. 6} \)
14. \( \mathcal{L}_{\text{adv}} \leftarrow \text{adversarial loss from} \mathcal{D} \quad \triangleright \text{Eq. 8} \)
15. \( \mathcal{L}_{\text{off}, \text{der}} \leftarrow \text{offset and derivation loss} \quad \triangleright \text{Eq. 9} \)
16. \( \mathcal{L}_{\text{all}} \leftarrow \text{overall training loss} \quad \triangleright \text{Eq. 10} \)
17. Update \( G \) by minimizing \( \mathcal{L}_{\text{all}} \)
18. Update \( D \) by maximizing \( \mathcal{L}_{\text{all}} \)
21. end while
we adopt Frechet Inception Distance (FID) (Heusel et al., 2017) and Depth (Ranftl et al., 2020) to calculate visual and geometry similarity, compared to the ground-truth image. We treat Percentage of Correct Keypoints (PCK@0.5) (Andriluka et al., 2014) as the correctness of the generated pose. To investigate the textual relevance of T3H results, we follow CLIP Score (CLIP-S) (Hessel et al., 2021) for the text-visual similarity. We fine-tune CLIP (Radford et al., 2021) on DeepFashion for a more accurate alignment in this specific domain. To have the fine-grained evaluation, we train a fashion classifier on DeepFashion labels and assess Fashion Accuracy (FA) of the generated human.

### Baselines
As a new task, we consider the following methods as the compared baselines.

- **Latent-NeRF** (Metzer et al., 2023) brings NeRF to the latent space and guides its generation by the given object and a text-to-image prior.
- **TEXTure** (Richardson et al., 2023) predicts the target human image, and Text2Human (Jiang et al., 2022) predicts the target human image, and treat Texformer to further build its 3D model.
- **CLIP-O** is inspired by AvatarCLIP (Hong et al., 2022), which customizes a human avatar from the description with CLIP text-visual alignment. We apply the guided loss to optimize a pre-trained EVA3D (Hong et al., 2023) for faster inference.
- **Texformer** (Xu and Loy, 2021) estimates the human texture from an image. Text2Human (Jiang et al., 2022) predicts the target human image, and we treat Texformer to further build its 3D model. For a fair comparison, all baselines are re-trained on face-blurred datasets and cannot produce identifiable human faces.

### Implementation Detail
We divide a human body into 16 parts and deploy individual StyleSDF (Or-El et al., 2022) for each volume rendering, and two following MLPs then estimate SDF and RGB values. We adopt the same discriminator as StyleSDF over fashion maps to distinguish between fake rendered humans and real images. We sample $N=28$ points for each ray and set $(m, n)$ to $(4, 8)$ for mix-

<table>
<thead>
<tr>
<th>Method</th>
<th>DeepFashion</th>
<th>SHHQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latent-NeRF</td>
<td>69.654</td>
<td>72.256</td>
</tr>
<tr>
<td>TEXTure</td>
<td>37.058</td>
<td>48.618</td>
</tr>
<tr>
<td>CLIP-O</td>
<td>25.488</td>
<td>34.212</td>
</tr>
<tr>
<td>CCH</td>
<td>21.136</td>
<td>32.858</td>
</tr>
</tbody>
</table>

Table 1: Overall results of pose-guided T3H.
4.2 Quantitative Results

Table 1 shows the pose-guided T3H results on DeepFashion and SHHQ, where we feed the estimated human mesh as the input object into Latent-NeRF and TEXTure. Although Latent-NeRF can portray body shapes in multiple angles from its latent NeRF space, the rendering is clearly counterfeit (higher FID and Depth). For TEXTure, the human architecture is constructed well by the given mesh (higher PCK). However, the estimated texture is still spatially inconsistent and contains inevitable artifacts (still higher FID). From the semantic aspect, Latent-NeRF and TEXTure borrow those trained diffusion models and depict the assigned appearance in the description (higher CLIP-S than CLIP-O). CLIP-O relies on EVA3D to produce feasible 3D humans (lower FID). While the external CLIP loss attempts to guide the fashion, the global alignment is insufficient to demonstrate detailed patterns (lower FA). Without those above drawbacks, our CCH learns to extract fashion semantics along with the compositional human generation, leading to comprehensive superiority across all metrics.

A similar trend can be found on SHHQ. Latent-NeRF and TEXTure exhibit related fashion patterns but are hard to present realistic humans (higher FID and Depth). CLIP-O produces a sharp human body with the correct pose, but not the assigned fashion (lower CLIP-S and FA) by the inexplicit alignment from CLIP. Table 2 presents the pose-free results. With the guided 2D image, Texformer contains the assigned clothing in the text (higher FA than CLIP-O). But the 3D reconstruction is unable to handle spatial rendering, resulting in low-quality humans (higher FID). With cross-modal attention and semantic discrimination, CCH exceeds baselines in both visual and textual relevance, making concrete human rendering with the corresponding fashion.

4.3 Ablation Study

We study each component effect of CCH in Table 3. Without the guided description, the model lacks the target fashion and results in a poor FA. This further highlights the importance of textual guidance for controllable human generation. When applying the traditional training (Reed et al., 2016), conditional GAN is insufficient to extract fashion semantics for effective T3H (not good enough CLIP-S). On the other hand, our cross-modal attention constructs a better fusion between fashion patterns and volume rendering, facilitating a significant improvement in depicting the desired human appearance. Moreover, semantic discrimination benefits fine-grained alignment and leads to comprehensive advancement.

Fine-tune CLIP-S as Evaluator. CLIP has shown promising text-visual alignment, which can calculate feature similarity between the generated human and the given text as CLIP-S (Hessel et al., 2021). Since our T3H is in a specific fashion domain, we consider the larger-scaled trained checkpoint from OpenCLIP (Ilharco et al., 2021) and fine-tune it as
Table 6: Time and GPU cost to perform T3H.

<table>
<thead>
<tr>
<th>Method</th>
<th>Time (sec)</th>
<th>GPU (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latent-NeRF</td>
<td>755.7</td>
<td>11250</td>
</tr>
<tr>
<td>TEXTure</td>
<td>103.7</td>
<td>12530</td>
</tr>
<tr>
<td>CLIP-O</td>
<td>181.6</td>
<td>15988</td>
</tr>
<tr>
<td>CCH</td>
<td>0.372</td>
<td>6258</td>
</tr>
</tbody>
</table>

Human Evaluation. Apart from automatic metrics, we conduct the human evaluation with aspects of 3D quality and fashion relevance. We randomly sample 75 T3H results and consider MTurk\(^2\) to rank over baselines and our CCH. To avoid the potential ranking bias, we hire 3 MTurkers for each example. Table 5 shows the mean ranking score (from 1 to 4, the higher is the better). CLIP-O and CCH are built upon EVAD3, which provides an articulate human body for superior 3D quality. Even if Latent-NeRF and TEXTure take pre-trained diffusion models to acquire visual guidance, CCH exhibits more corresponding fashion via cross-modal fusion. This performance trend is similar to our evaluation, which supports the usage of CLIP-S and FA as metrics.

Inference Efficiency. In addition to T3H quality, our CCH also contains a higher efficiency. Table 6 shows the inference time and GPU cost on a single NVIDIA TITAN RTX. All baselines take more than 100 seconds since they require multiple iterations to optimize the 3D model from an external alignment. In contrast, we extract fashion semantics and carry out T3H in one shot. Without updating the model, we save the most GPU memory. In summary, CCH surpasses baselines on both quality and efficiency, leading to an effective and practical T3H.

4.4 Qualitative Results.

We demonstrate the qualitative comparison of pose-guided T3H in Fig. 3. Although Latent-NeRF can portray the 3D human based on the given mesh, it only presents inauthentic rendering. TEXTure generates concrete humans, but there are still obvious cracks and inconsistent textures from different angles of view. Moreover, both of them fail to capture “three-point”, where the rendered lower clothing is

\(^2\)Amazon MTurk: https://www.mturk.com
Figure 5: Qualitative examples of animatable T3H, where the motion is also controlled by the text.

Figure 6: Qualitative examples of pose-control T3H.

Figure 7: Qualitative examples of our fashion classifier, which provides fine-grained labels for real/fake humans.

incorrectly depicted as long pants. Because CLIP provides an overall but inexplicit alignment to the description, CLIP-O is limited and exhibits vague “denim” or “long-sleeved”. This observation further indicates the flaw of CLIP in detailed fashion patterns, even if it has been fine-tuned on the target dataset. In contrast, our CCH adopts cross-modal attention with NeRF, contributing to high-quality T3H with fine-grained fashion controllability. Fig. 4 shows the pose-free results. Texformer relies on a 2D image to estimate its 3D texture. Despite containing the assigned fashion, it is still restricted by the capability of 3D reconstruction, resulting in a low-resolution rendering. By learning text-to-3D directly, CCH can produce textual-related humans from random poses with clear visual patterns.

Pose-control T3H. Since our CCH is generating 3D humans from given SMPL parameters, as illustrated in Fig. 6, we can further control T3H with a specific pose. Different fashion descriptions make a human body present diverse appearances; different poses then guide the character to express rich body language. This flexibility in controlling appearance and pose allows for better practical customization.

Animatable T3H. In addition to static poses, CCH can benefit from dynamic motions to achieve animatable T3H. Fig. 5 adopts MotionDiffuse (Zhang et al., 2022) to create the assigned action also from the text and apply it to our produced 3D models. In this way, we prompt them to “raise arms” or “walk” for favorable dynamic scenarios.

5 Conclusion

We present text-guided 3D human generation (T3H) to create a 3D human by a fashion description. To learn this from 2D collections, we introduce Compositional Cross-modal Human (CCH). With cross-modal attention, CCH fuses compositional human rendering and textual semantics to build a concrete body architecture with the corresponding fashion. Experiments across various fashion attributes show that CCH effectively carries out T3H with high efficiency. We believe T3H helps advance a new field toward vision-and-language research.

Ethics Discussion and Limitation. Our work enhances the controllability of 3D human generation. To prevent identity leakage, we blur out faces prior to training and avoid risks similar to DeepFake (Korshunov and Marcel, 2018). Because we depend on SMPL parameters, an inaccurate estimation causes...
a distribution shift and quality degradation. For the datasets, they reveal narrow viewing angles, which results in visible artifacts of 3D consistency.

Acknowledgements. We appreciate the anonymous reviewers for constructive feedback. The research presented in this work was funded by Meta AI. The views expressed are those of the authors and do not reflect the official policy or position of the funding agency.

References


In Special Interest Group on Computer Graphics and Interactive Techniques in Asia (SIGGRAPH Asia).


Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and


Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. 2022a. CLIP-NeRF: Text-and-Image Driven Manipulation of Neural Radiance Fields. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.


