

TeRA: Rethinking Text-driven Realistic 3D Avatar Generation

Supplementary Material

1. Dataset Details and Comparisons

A sample of our dataset is detailed in Fig. 1. And as shown in Tab. 1, our dataset is the largest text-annotated multi-view human dataset to date, containing significantly more identities than MVHumanNet—the only comparable dataset with textual annotations. Moreover, our annotations are more accurate and comprehensive than previous datasets.

Dataset	Frames	ID	View	Text Caption
HuMMan	60M	1000	10	55
HUMBI	26M	772	107	55
DNA-Rendering	67.5M	500	60	55
MVHumanNet	645.1M	4500	48	51
THuman2.1	-	2500	Free	55
2K2K	-	2050	Free	55
Ours	2.4M	100K	24	51

Table 1. Comparisons of datasets.

2. Data Caption

Since it is challenging for a large language model to output accurate labels of varying lengths in a single conversation, we annotate text over three rounds of dialogue. Precisely, to capture as much information as possible from the input images, we prompt the Qwen-2.5VL model in the first round to provide detailed descriptions of four orthogonal views of the human body, including various pertinent details. In this round, Qwen-2.5VL outputs a comprehensive descrip-

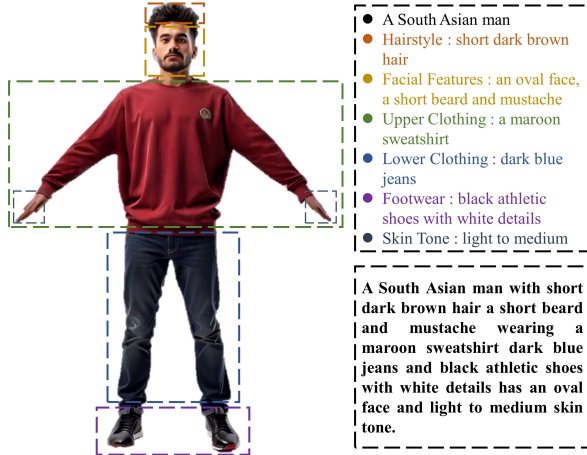
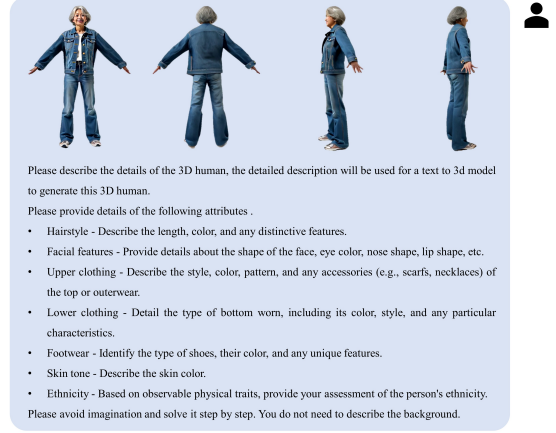


Figure 1. Details of our dataset



Certainly! Here is a detailed description of the 3D human based on the provided images:

- Hairstyle - **Length**: The hairstyle is short, reaching just above the earlobes. **Color**: The hair is gray. **Distinctive Features**: The hair appears to be styled in a neat, slightly wavy manner.
- Facial Features - **Shape of Face**: The face appears to be oval-shaped. **Eye Color**: The eye color is not clearly visible due to the angle and lighting. **Nose Shape**: The nose appears to be medium-sized with a straight bridge. **Lip Shape**: The lips are closed, and they appear to be of average thickness.
- Upper Clothing - **Style**: The person is wearing a denim jacket. The jacket has a classic design with long sleeves, a collar, and button closures. **Color**: The denim jacket is blue. **Pattern**: There are no visible patterns on the jacket; it is solid blue. **Accessories**: There are no visible accessories such as scarves or necklaces.
- Lower Clothing - **Type**: The person is wearing jeans. **Color**: The jeans are blue, matching the jacket. **Style**: The jeans have a straight-leg style with visible stitching along the seams. **Characteristics**: There are no distinctive features other than the standard design of jeans.
- Footwear - **Type**: The person is wearing sneakers. **Color**: The sneakers are primarily white with some black details. **Unique Features**: The sneakers have a classic design with laces and a rubber sole.
- Skin Tone - **Description**: The skin tone is light.
- Ethnicity - **Assessment**: Based on observable physical traits, the person appears to be of East Asian descent. This assessment is made based on the facial structure and skin tone.

This detailed description should help in generating a similar 3D human model.

This is a hard problem. Carefully summarize in ONE caption aiming for **no more than 40 words** based on the captions (possibly incorrect) you generate before. The caption will be used for a text to 3D model to generate this 3D human. Ensure the summary is concise and captures the essential information without including any additional commentary or unnecessary details. Please avoid hallucination. Please directly output the detailed caption without extra text. Provide the caption in a simple, plain text format with no markdown, bullet points, or special formatting.

A short-haired East Asian woman with a gray hairstyle wears a blue denim jacket and jeans, white sneakers, and has light skin tone. Her face is oval-shaped with a straight nose and average lips.

This is a hard problem. Carefully summarize the given caption describing the person. The caption will be used for a text to 3D model to generate this person. Ensure the summary is concise and captures the essential information without including any additional commentary or unnecessary details. Please avoid hallucination. Please give me FIVE versions of the caption. ONE caption per line. The captions should be more and more brief. The first caption should be the most detailed and the last caption should be the most brief. The first caption should be less than 16 words and the last caption should describe ALL major feature about the person AS SIMPLY AS POSSIBLE for at most 8 words. Please directly output the caption without extra text. Provide the caption in a simple, plain text format with no markdown, bullet points, numbering, or special formatting.

A short-haired East Asian woman with gray hair wears a blue denim jacket, jeans, and white sneakers.
An East Asian woman with gray hair wears a denim jacket, jeans, and sneakers.
A woman with gray hair wears a blue jacket, jeans, and white shoes.
A gray-haired woman in a jacket, jeans, and sneakers.
Gray-haired woman, jacket, jeans, sneakers.

Figure 2. An example of our captioning process.

tion; however, to meet the input text length requirements of the CLIP model, these verbose descriptions must be condensed. In the second round, we feed the output from the first round into the Qwen2.5 model and request that it summarize the detailed description into a long annotation of no more than 40 words. Finally, to further enrich the textual content, we ask Qwen2.5 to further condense the outputs from the first two rounds into five descriptions of varying lengths, with the longest being no more than 16 words and the shortest containing at least 8 words. During training, one of these five short descriptions or the 40-word long annotation is randomly selected as the condition. The complete text annotation process is illustrated in Fig. 2.

3. Architecture of Distillation Decoder

The distillation decoder consists of a UV code decoder and a Gaussian attribute decoding head. The UV code decoder includes two transposed convolution layers and two convolution layers, with an input feature size of $256 \times 256 \times 32$ and an output code size of $1024 \times 1024 \times 32$. The output code from the UV encoder-decoder is split into two parts: the first 16 channels are regarded as the geometry code, and the remaining 16 channels as the texture code. The Gaussian attribute decoding head consists of three convolutional heads, each comprising two or three convolutional layers. These Gaussian attribute decoding heads are responsible for decoding the geometry code and texture code into five 3D Gaussian attributes in the SMPL-X texture space.

than TeRA. Our representation simplifies the learning of the target distribution and enables downstream applications.

5. More Results

We show more renderings of our generated models with input text description in Fig. 4 and Fig. 5.



Figure 3. Additional qualitative comparisons with general 3D methods.

4. More Comparisons

As far as we know, SDS-based models are the only available text-to-3D-avatar methods. We have evaluated additional baselines of general 3D reconstruction methods, including LGM (text \rightarrow multi-view \rightarrow 3DGS), GVGen (direct 3D), and DiffSplat (2D-diffusion \rightarrow 3D) in Fig. 3; all deliver lower visual fidelity and weaker prompt adherence

A Caucasian man with short gray hair and a light complexion wears a brown sweater, navy blue jeans, and brown shoes



A Caucasian woman with short gray hair, an oval face, thin lips, and fair skin wears a brown leather jacket, matching wide-legged pants, and brown shoes



A Hispanic man with short black hair, and medium skin tone wears a green t-shirt, denim jeans, and brown boots



A man with a shaved head and dark skin wears a yellow t-shirt, denim shorts, and black sandals



Figure 4. More results of text-guided generation

A man with short gray hair and a beard wears a brown jacket, white henley shirt, and dark jeans



A woman with long straight hair in a high ponytail wears an olive green military jumpsuit with cargo pants and black combat boots, light to medium skin tone



A woman with short curly hair and glasses long red skirt, and black flats



A woman with straight brown hair and light skin wears a navy blue wrap dress and black flats



Figure 5. More results of text-guided generation

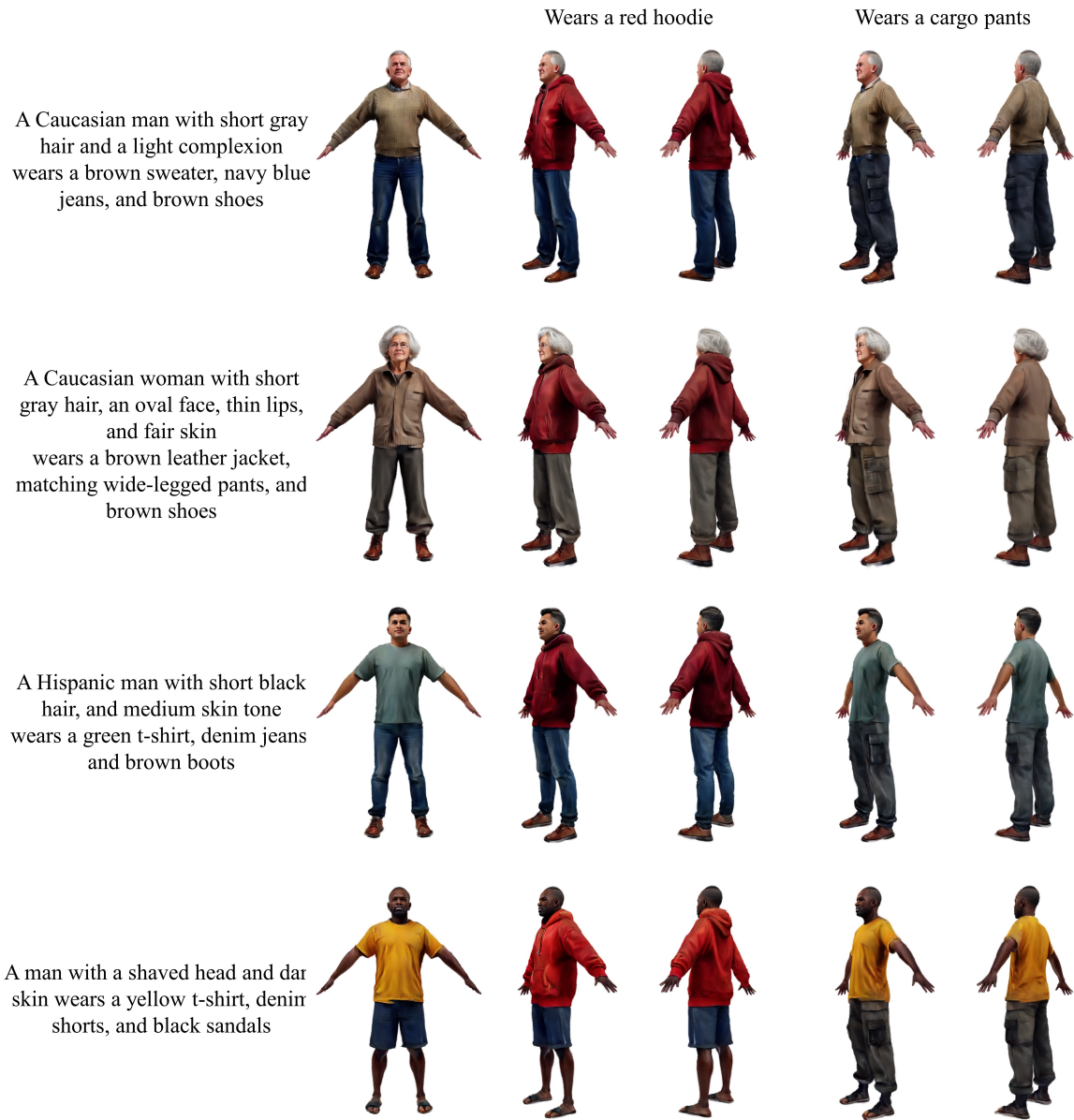


Figure 6. More results of text-guided virtual try-on