HumanDiffusion: a Coarse-to-Fine Alignment Diffusion Framework for Controllable Text-Driven Person Image Generation

Kaiduo Zhang$^{1,2}$, Muyi Sun$^1$, Jianxin Sun$^{1,2}$, Binghao Zhao$^{1,2}$, Kunbo Zhang$^1$, Zhenan Sun$^{1,2}$, Tieniu Tan$^{1,2,3}$

$^1$ Center for Research on Intelligent Perception and Computing, NLPR, CASIA
$^2$ School of AI, University of Chinese Academy of Sciences (UCAS)
$^3$ Nanjing University (NJU)

{kaiduo.zhang, muyi.sun, jianxin.sun, binghao.zhao}@cripac.ia.ac.cn, kunbo.zhang@ia.ac.cn, {znsun, tnt}@nlpr.ia.ac.cn

Abstract

Text-driven person image generation is an emerging and challenging task in cross-modality image generation. Controllable person image generation promotes a wide range of applications such as digital human interaction and virtual try-on. However, previous methods mostly employ single-modality information as the prior condition (e.g. pose-guided person image generation) or utilize the preset words for text-driven human synthesis. Introducing a sentence composed of free words with an editable semantic pose map to describe person appearance is a more user-friendly way. In this paper, we propose HumanDiffusion, a coarse-to-fine alignment diffusion framework, for text-driven person image generation. Specifically, two collaborative modules are proposed, the Stylized Memory Retrieval (SMR) module for fine-grained feature distillation in data processing and the Multi-scale Cross-modality Alignment (MCA) module for coarse-to-fine feature alignment in diffusion. These two modules guarantee the alignment quality of the text and image, from image-level to feature-level, from low-resolution to high-resolution. As a result, HumanDiffusion realizes open-vocabulary person image generation with desired semantic poses. Extensive experiments conducted on DeepFashion demonstrate the superiority of our method compared with previous approaches. Moreover, better results could be obtained for complicated person images with various details and uncommon poses.

1. Introduction

High-fidelity and high-diversity person image generation is an emerging and challenging task in multi-modal interaction systems. Previous approaches [Sarkar et al., 2021, Ren et al., 2022, Cheong et al., 2022, Zhu et al., 2019, Zhang and Zhao, 2021, Siarohin et al., 2018] mostly focus on single-modality guided generation (e.g. skeleton-based or semantic-guided). Recently, with the rapid development of the cross-modality image generation [Sun et al., 2022, Ho and Salimans, 2022, Gu et al., 2022, Gafni et al., 2022, Zhou et al., 2021, Kim and Kwon, 2022, Ramesh et al., 2022], some researchers try to explore the text-guided human image synthesis with the preset vocabularies [Jiang et al., 2022b, Liu et al., 2022, Xu and Chen, 2021]. However, their methods are limited in practical applications. More flexible and free-style text-driven person image generation is desirable to be explored.

In general, human synthesis methods are mostly implemented with traditional generation frameworks, such as Generative Adversarial Networks (GANs) [Sarkar et al., 2021, He et al., 2022, Lewis et al., 2021, Fu et al., 2022] or Variational Auto-Encoders (VAEs) [Jiang et al., 2022b]. From the perspective of the target tasks, the human synthesis can be divided into two categories: human pose trans-
fer [Ren et al., 2022, Zhu et al., 2019, Zhang and Zhao, 2021, Zhang et al., 2021a] and text-driven person image generation [Xu and Chen, 2021, Jiang et al., 2022b, Liu et al., 2022]. Pose transfer aims to transfer the human appearances from the source image into the desired pose, and commonly follows the GANs-based architecture. The main issue of human pose transfer is to align the cross-domain image features [Zhang et al., 2021a]. Though the generation quality of these methods is satisfactory, they are not intuitive and user-friendly in real scenes. In contrast, as a carrier that can express information more flexibly, the text shows more interactivity and controllability for person image generation. The rich expressions in the text bring more diversity to the human body in colors, textures, and even clothing types as shown in Figure 1.

Recent text-driven person image generation methods have realized basic human image generation from input prompts which consisted of preset words [Liu et al., 2022, Xu and Chen, 2021, Jiang et al., 2022b]. Verbal-Person [Liu et al., 2022], a GANs-based network, introduces the local text-relevant and global full-body discriminators to promote the text correspondence with the target skeleton under the supervision of the paired text annotations. Similarly, Image-Text Shared Space (ITSS) [Xu and Chen, 2021] proposes a flexible and integrated GANs-based framework for text-guided human image manipulation, which explores different moving distances in the shared latent space of the text and image. Different from the low-quality generation methods above mentioned [Liu et al., 2022, Xu and Chen, 2021], Text2Human [Jiang et al., 2022b] designs a multi-stage framework for high-resolution text-driven human synthesis which utilizes a texture-aware codebook to realize fine-grained feature representation. However, these methods mostly use preset words as the supervision and show limitations in processing multi-modality feature fusion and alignment, which are not flexible for complex human image description and generation. Therefore, there is an urgent demand to introduce an open-vocabulary method for free-style human appearance description and appropriate frameworks for multi-modality fusion.

Recently, diffusion models such as denoising diffusion probabilistic models (DDPM) [Ho et al., 2020] and score-based generative models [Song and Ermon, 2019, Song et al., 2020b] show powerful capability in image generation tasks [Nichol et al., 2021, Gafni et al., 2022, Ramesh et al., 2022]. The latest researches [Dhariwal and Nichol, 2021, Ramesh et al., 2022, Kawar et al., 2022] perform better potential in text information perception and high-quality image generation, compared to VAEs [Doersch, 2016] and GANs [Goodfellow et al., 2020]. These days, diffusion models have attracted wide attention and have been applied to various fields, especially in cross-modality generation tasks.

For flexible and free-style person image generation, we propose HumanDiffusion, a newly designed diffusion-based framework with coarse-to-fine cross-modality feature alignment modules. In this coarse-to-fine framework, two collaborative modules are designed for high-quality image generation, a Stylized Memory Retrieval (SMR) module and a Multi-scale Cross-modality Alignment (MCA) module. The SMR module is located before the input appearance features, which introduces a disentangled memory network to refresh the semantic representations from low-quality information into fine-grained features and enhance the text features with the image-aware representations for sampling. Thus, the SMR module endows HumanDiffusion with the ability of text-free training and text-driven sampling. The MCA module is designed with the cascaded cross-attention mechanism and integrated into the diffusion backbone, which gradually aligns the middle-layer image-aware features and input text-aware features from low-resolution to high-resolution. Furthermore, the HumanDiffusion is created with a compact structure and implemented with an end-to-end text-free training strategy.

The contributions are summarized as follows:

- A coarse-to-fine alignment diffusion framework, named HumanDiffusion, is proposed for flexible, free-style, and text-driven person image generation.
- A Multi-scale Cross-modality Alignment (MCA) module is designed into the diffusion framework, which gradually improves the feature alignment degrees of the text and image, from low-resolution to high-resolution.
- A Stylized Memory Retrieval (SMR) module is proposed to provide more fine-grained image-aware partial features, which also enhances the open-vocabulary capability of text guidance.
- Extensive experiments conducted on the DeepFashion dataset demonstrate the superiority of the HumanDiffusion. Better results are obtained for complicated person generation and manipulation.

2. Related Works

2.1. Controllable Person Image Generation

Controllable person image generation is a challenging vision task [Sarkar et al., 2021, He et al., 2022, Fu et al., 2022, Jiang et al., 2022b, Lewis et al., 2021, Ma et al., 2018, Li et al., 2019, Wang et al., 2022b]. Researchers try to apply different generative models to handle this problem. Most existing models synthesize human images in two aspects, pose generation [Zhang and Yang, 2022, Zhu et al., 2019, Zhang and Zhao, 2021, Zhang et al., 2021a] and texture refinement [Lewis et al., 2021, He et al., 2022]. SAWN
Figure 2. Overview of the HumanDiffusion framework. In the training procedure, the Stylized Memory Retrieval (SMR) module is designed to obtain fine-grained image-aware features which support text-free training; a diffusion-based network is proposed for cross-modality feature alignment with the help of the Multi-scale Cross-modality Alignment (MCA) module under the pose condition. In the sampling procedure, given a text description and a semantic pose map, a text-driven person image could be generated with HumanDiffusion.

[Zhang et al., 2021b] uses a spatially-adaptive warped normalization to align the style and pose features. Pose-with-Style [Albahar et al., 2021] proposes a pose-conditioned StyleGAN [Fu et al., 2022] framework and wraps the target pose with the source image details. TryOnGAN [Lewis et al., 2021] proposes a semantic-driven model to synthesize human images for virtual try-on. However, these methods are not flexible, and can not be applied interactively in real scenes. Text2Human [Jiang et al., 2022b] introduces a text-driven perspective for human synthesis. The Verbal-Person [Liu et al., 2022] employs a structure-enhanced GAN model to generate similar text-guided human images. However, these GANs-based methods are limited in cross-modality feature alignment, and commonly need paired text annotations for training. To overcome these problems, we propose an end-to-end and text-free training framework to generate person images, conditioned on the target semantic poses and natural language descriptions.

2.2. Text to Image Generation with Diffusion

Recently, text-to-image generation tasks have made great progress with the emergence of Diffusion models [Nichol et al., 2021,Gafni et al., 2022,Ramesh et al., 2022,Gu et al., 2022,Liu et al., 2021b,Ramesh et al., 2021]. Due to the ability to preserve the semantic structure of data, diffusion models can perform higher diversity and fidelity than previous generation frameworks [Kim and Kwon, 2022]. Many researches have explored text-conditional image generation, in which the most outstanding schemes should be the huge cross-modality models like DALLE-2 [Ramesh et al., 2022], GLIDE [Nichol et al., 2021]. These models can generate amazing images according to the prompts people desire, which can even synthesize scenes that people have never seen. Furthermore, Retrieval-Diffusion [Blattmann et al., 2022] introduces the nearest neighbor retrieval to improve the generation quality and diversity. Imagic [Kawar et al., 2022] performs better results for diverse text-driven image editing. However, these models are all general text-to-image frameworks on natural image generation which can not perceive fine-grained features for specific downstream tasks such as person image generation. In this paper, we focus on the sub-task of text-driven person image generation and design a specific diffusion framework, which could deeply explore the human image properties (e.g. semantic structural information).

2.3. Memory Networks

The idea of memory networks is mainly inherited from RNN, including LSTM [Hochreiter and Schmidhuber, 1997] and GRU [Cho et al., 2014]. The memory networks have been applied to various vision tasks, including image captioning [Chunseong Park et al., 2017], text-to-image synthesis [Zhu and Pan, 2019]. Weston et al. [Miao and Wei, 2020] propose memory networks to overcome the limitations of RNNs. Zhu et al. [Zhu and Pan, 2019] utilize a multi-modal memory module to refresh and clear blurred images for text-to-image generation. Recently, retrieval modules appear in text-to-image diffu-
sion models such as KNN-Diffusion [Sheynin et al., 2022] and Retrieval-Diffusion [Blattmann et al., 2022], which retrieve the nearest neighbors of the condition images. These neighbors provide extra details to enhance the image expression. However, they are mostly fixed and preset in the retrieval module, which limits the representative ability of fine-grained features. Inspired by the memory mechanism, we design a dynamic memory module to provide text-aware features which can be more consistent with the distribution of the training data, and supply better additional detailed information which enhances clothes representation in free-style.

3. Method

HumanDiffusion is implemented on the diffusion-based U-Net framework as shown in Figure 2, which consists of two parts: the Stylized Memory Retrieval (SMR) module and the Multi-scale Cross-modality Alignment (MCA) module. In the following, we will introduce this method in detail with the training and sampling procedure specifically.

As for the training process, first given a human image and the corresponding semantic map $S$, the coarse features $F_{p1}$ are obtained by the CLIP image encoder. The $F_{p1}$ can be also treated as the text-aware representations, which is attributed to the powerful alignment ability of CLIP. Thus, text-free training could be achieved. Then, $F_{p1}$ is sent into SMR to update the memory blocks and retrieve the fine-grained appearance features $F_{p2}$, which is image-aware and combined with $t$ to get the global condition $t'$. $t'$ is then injected into SPADE ResBlocks (SRES) combined in HumanDiffusion. Meanwhile, the semantic map $S$ is spatially disentangled into $seg$ and injected into SRES as the pose condition. $F_{p1}$ is then transferred into a Self-Attentive Alignment module (SAA) to obtain the highlight features $F_{ps}$. Then $F_{ps}$ is sent into MCA. The MCA module performs Cross-modality Attentive Alignment (CAA) between $F_{ps}$ and the middle image features $f_i$ in SRES, from low-resolution to high-resolution. Finally, the refined human features are obtained for finer image generation.

For the sampling process, the desired pose map and the text description are treated as inputs. Then a satisfying human image could be obtained. In the following, we will describe the significant modules in detail.

3.1. Stylized Memory Retrieval

To generate high-quality person images, appearance features act as the most important prior. To obtain fine-grained appearance features as the global condition of the diffusion, we propose the Stylized Memory Retrieval (SMR) module in data processing, as shown in Figure 3. We design disentangled memory blocks to store the image-aware CLIP features [Radford et al., 2021], which corresponds to the specific human parts. There are two steps to establish this memory module, initialization with external databases and updating with the training dataset. As shown in Figure 3, the SMR is first initialized with the CLIP features of clothing images in free-style. Then, the memory blocks are updated with the training data (i.e. image CLIP features). With the external prototype clothing features and the specific characteristics from the training set, the SMR could provide fine-grained and image-aware representations of the input appearances.

**Memory Initialization.** We establish the disentangled memory blocks for different semantic categories to store the semantic-aware features, where $n$ represents the number of disentangled human parts, and $c$ is the feature dimension. An external database [Liu et al., 2016] is treated as the retrieval source, which contains clothes images in various styles. 500 images are selected as the memory slots for each category, which almost contain all clothing styles. Thus, the memory has been established with $n$ blocks ($M \in \mathbb{R}^{n \times (m \times c)}$) and each block ($E \in \mathbb{R}^{m \times c}$) contains $m$ memory slots $e \in \mathbb{R}^c$. The restored memory slots are encoded by the CLIP encoder.

**Memory Updating.** In the training procedure, the memory blocks are fine-tuned with the DeepFashion training dataset gradually. Given a human image, the CLIP encoder maps the image into latent representations $F \in \mathbb{R}^{n \times c}$ ($F_{p1}$), which represents the disentangled parts of this human. We use the similarity between the input latent feature and the relevant memory slots to update the memory blocks. For $ith$ semantic in the image, we compute the cosine distance $s_{ij}$, between $F_i$ and the $jth$ memory slot $e_{ij}$, which is formulated as follows:

$$s_{ij} = \frac{e_{ij} F_i^T}{\|e_{ij}\| \|F_i\|}$$  \hspace{1cm} (1)

Then, we seek the memory slot $e_{ik}$, which is most relevant with the input query $F_i$:

$$k = \text{argmax}_j (s_{ij})$$  \hspace{1cm} (2)
To update the relevance between the memory with the training images, we use the latent vector $F_i$ in each batch, to update the memory slot $e_{ik}$, which also represents the initial features for the query $F_i$. The memory updating process can be formulated as follows:

$$e_{ik} \leftarrow \beta e_{ik} + (1 - \beta)F_i$$

(3)

where $\beta \in [0, 1]$ is a decay rate set to 0.995 in our paper. Thus, the memory blocks integrate the prototype cloth features and the specific characteristics of the training set.

**Memory Reading.** After the memory updating, we refresh the input latent vector $M_i$ to a fine-grained representation $\tilde{M}_i$. We employ an attentive aggregation training strategy. To begin with, the cosine similarity matrix $\gamma = \{\gamma_{ij} | i = 1, \ldots, n, j = 1, \ldots, m\}$ is computed through Equation 1 again. Then, the soft scores $A = \{a_{ij} | i = 1, \ldots, n, j = 1, \ldots, m\}$ are formulated by a softmax operation.

$$a_{ij} = \frac{\exp(\gamma_{ij})}{\sum_{m=1}^{m} \exp(\gamma_{ij})}$$

(4)

Finally, the memory-based latent vectors $\tilde{F}_i (F_{p2})$ are constructed by aggregating memory slots with soft scores.

$$\tilde{F}_i = \sum_{j} a_{ij} e_{ij}$$

(5)

To summarize, we propose a Stylized Memory Retrieval Module which utilizes an external database to get fine-grained image-aware feature vectors $F_{p2}$ which provides abundant clothing representations in free-style. The SMR module is jointly trained with the diffusion networks in the following. Furthermore, while sampling, the retrieval operation helps to refresh the text CLIP features to be image-aware, which provides the potential for text-free training.

**3.2. HumanDiffusion: Training**

In this section, we will introduce the training procedure of the HumanDiffusion framework specifically.

**Data Processing.** For each training iteration, we first perform the dot multiplication on the input image $I$ and the corresponding semantic map $S$ to $n$ parts $I_i$ (specific clothes appearances). Then, all the $I_i \in I$ is sent into the CLIP image encoder to extract a $n \times 512$ features:

$$I_i = I \otimes S_i$$

(6)

$$F_{p1i} = CLIP(I_i)$$

(7)

where $S_i$ denotes one-part semantic map in human images, $F_{p1i}$ is the partial CLIP features in $F_{p1}$. Then the $F_{p1}$ is sent into the SMR module to get refreshed features $F_{p2}$ which are fused with the time step $t$ to act as the global condition $t'$. The $seg$ from semantic map $S$ is acted as the pose condition and is also injected into the SPADE ResBlocks.

**SPADE ResBlocks.** Inspired by [Wang et al., 2022a, Park and Liu, 2019], we design the SPADE ResBlocks (SRES) to inject the pose semantic information $seg$ and the global condition $t'$ for human pose control and image-aware feature guidance. As shown in Figure 4, the SRES module employs spatially adaptive normalization (SPADE) [Park and Liu, 2019] to inject the pose feature $seg$.

$$f_{spade} = SPADE(f_{in}, seg)$$

(8)

$$f_{t'} = CONV(ReLU(f_{spade})) \times (1 + \omega) + b$$

(9)

As shown in Equation 8, $f_{in}$ is the input features of SRES, and $f_{spade}$ is the output of SPADE which contains more pose information. $t'$ is injected as shown in Equation 9, in which $\omega$ and $b$ are modulation coefficients obtained by sending the global condition $t'$ into Fully Connected (FC) layers. Then $f_{t'}$ contains more fine-grained appearance features. In summary, the SRES module helps the diffusion model to obtain more pose and image-aware feature information.

**Multi-scale Cross-modality Alignment.** Previous general text-to-image researches like Dalle2 [Ramesh et al., 2022] and GLIDE [Nichol et al., 2021] mostly utilize the powerful coding capabilities of the CLIP to measure and align the distance between the output images and the input prompts. For human image generation, human images have strong structural information. It is difficult to generate image details of complex structures depending on simple alignment strategies or constraints. As shown in Figure 5, we design a Cross-modality Attentive Alignment (CAA) module which
constitutes MCA in the multi-scale resolutions of the diffusion backbone, to gradually align the fine-grained partial features $F_{p1}$ with the U-Net middle features $f_i$. In this module, we treat $F_{p1}$ as the text-aware encoding and $F_{p2}$ as the image-aware encoding.

Before the MCA, for distilling finer features that highlight the important parts of $F_{p1}$, we perform self-attention to obtain fine-grained partial features $F_{ps}$. Then, the MCA module will further align the cross-modality features from coarse to fine. CAA is carried out through multi-scale resolutions from $8 \times 8$ to $32 \times 32$. Specifically, as shown in Figure 5, CAA has three learnable projection matrices $W_Q$, $W_K$, $W_V$. The middle features $f_i$ are treated as queries $Q \in \mathbb{R}^{b \times r \times r}$, where $b$ is the batchsize, $r$ is the resolution of middle features. The fine-grained features $F_{ps}$ from SAA are performed as the keys $K$ and values $V$, which are queried by $Q$ for alignment information distillation:

$$Q_i = f_i W_Q, K = F_{ps} W_K, V = F_{ps} W_V$$

$$f_{align} = CAA(Q_i, K, V)$$

$$f_i = f_i + f_{align}$$

Note $f_{align}$ is the aligned features. $f_i$ is the output features that are sent into the next SRES.

After the coarse-to-fine alignment operations in Equations 10-12, the diffusion is more sensitive to spatial information, and each part of the features could obtain better correspondences with its text descriptions. As a result, the denoising diffusion procedure could be improved as Equation 13. In each step at time step $t$, $x_{t-1}$ and $x_t$ are the input and output image maps, $t^\prime$ and seg are the global and pose conditions, $F_p$ is the text-aware features, $\mu_d$ and $\Sigma_d$ are the mean and variance of the output which is decided by the U-Net parameter $\theta$.

$$p_\theta(x_{t-1} | x_t) = \mathcal{N}(x_{t-1}; \mu_d(x_t, (t^\prime, \text{seg}, F_p)),$$

$$\Sigma_d(x_t, (t^\prime, \text{seg}, F_p))$$

### 3.3. HumanDiffusion: Sampling

As shown in the bottom of Figure 2, for human image sampling, the target semantic map and the desired open-vocabulary text description is sent into the framework. First, the text description is disentangled according to the partial image structure and then encoded by the CLIP text encoder to get partial text features $F_{p3}$. Then SMR refreshes $F_{p3}$ and gets the fine-grained image-aware information, which promotes HumanDiffusion to use only images for training and open-vocabulary texts for sampling.

Similar to the training process, the initial features $(F_{p3})$, refined text features, and the semantic pose map are all input into HumanDiffusion to generate the target person.

Finally, a human image that conforms to the text description and appears in the target pose is obtained.

### 4. Experiments

#### Dataset.
We conduct our experiments on DeepFashion-MultiModal dataset [Liu et al., 2016] to verify the effectiveness of our method. As for the external memory database, we employ the In-shop Clothes Retrieval database [Liu et al., 2016], which contains large amounts of partial clothes images with clear detailed photographs. The DeepFashion MultiModal dataset contains 12701 full-body images with the corresponding semantic maps and text descriptions. The semantic maps contain parsing labels of 24 classes. We follow the setup of Text2Human [Jiang et al., 2022b] to split the training and testing set.

#### Implementation Details.
The diffusion model is implemented by PyTorch [Li et al., 2020] with 4 NVIDIA Tesla-A6000 GPUs under the batchsize 24. We use Adam optimizer [Kingma, 2014] as the gradient descent method, and set the learning rate $\alpha = 0.0001$. The model is trained for 240000 iterations and finetuned after 100000 iterations with the $\alpha$ set to 0.00001. We select 10 parts from the semantic maps, which can describe the basic human clothes appearances (background, top, outer, skirt, dress, pants, hair, face, shoes, skin). Similar to previous diffusion-based methods, we first train the model under $128 \times 256$ resolution and then train a super-resolution upsampling model up to $256 \times 512$.

#### Text-free Training.
HumanDiffusion can only use human images and semantic pose maps as inputs to train the model. With the help of SMR, whether the CLIP image features or text features $(F_{p1}/F_{p3})$ could be treated as text-aware embedding, whenever training or sampling.

#### Evaluation Metrics.
Three objective indicators and a user study are used to evaluate the generation performance. Frechet Inception Distance (FID) measures the fidelity of the generated images [Heusel et al., 2017]. Learned Perceptual Image Patch Similarity (LPIPS) computes the distance between the generated images and ground truths in the perceptual domain [Zhang et al., 2018]. Multi-Scale Structural Similarity (MS-SSIM) evaluates the structural similarity between two images [Wang and Simoncelli, 2003]. The user study is a subjective evaluation of the image quality and caption similarity.

#### 4.1. Text-Driven Person Image Generation

##### 4.1.1. Text-Driven Quantitative Results

As shown in Table 1, HumanDiffusion obtains competitive results with Text2Human in FID and outperforms it in LPIPS and MS-SSIM, where HumanDiffusion employs a more compact and flexible framework. Furthermore, the results from GLIDE and ablation experiments show that the SMR and MCA modules could improve human images in terms of image quality and structural similarity.
Figure 6. Quality Comparisons on image generation under the guidance of text on Text2Human, GLIDE, HumanDiffusion, and the ablations of HumanDiffusion without SMR and MCA. HumanDiffusion could express human appearance in open-vocabulary and free-style words. Compared to other methods, HumanDiffusion’s generation results are mostly corresponding to descriptions. Zoom in for better details.

Figure 7. Quality Comparisons on image generation under the guidance of semantic. The leftmost of each group is the input semantic map, columns from left to right represent the results of Pix2Pix, Spade, LatentDiffusion, Text2Human, and two sampled images by HumanDiffusion. HumanDiffusion can generate diverse human images in high-quality. Zoom in for better details.

Table 1. Quantitative comparisons and ablations on Text-Driven Person Image Generation with DeepFashion dataset.

| Methods         | FID↓ | LPIPS↓ | MS-SSIM↑ |
|-----------------|------|--------|---------|
| Text2Human      | 29.53| 0.2154 | 0.7358  |
| GLIDE           | 56.53| 0.3215 | 0.5658  |
| w/o SMR         | 38.43| 0.2224 | 0.7326  |
| w/o MCA         | 34.86| 0.2382 | 0.7022  |
| HumanDiffusion  | 30.42| 0.2095 | 0.7453  |

4.1.2 Text-Driven Qualitative Results

As shown in Figure 1 and 6, given a target pose and the appearance text description, the generation results are obtained for illustration. It is worth noting that Text2Human uses only preset words (e.g. pure color, denim, floral) to describe human appearances. HumanDiffusion could use free-style words (e.g. ‘shinning’, ‘colorful stripes’, ‘red floral’) to describe appearances, which can generate corresponding images in high-quality, as shown in Figure 1. For a fair comparison, we employ similar descriptions in Figure 6, which Text2Human uses to generate. HumanDiffusion could obtain comparable results with a better match with the text descriptions. Meanwhile, HumanDiffusion also performs better details and clearer disentanglement in different regions (e.g. generating high-fidelity hands, faces, and clear semantic boundaries), whereas Text2Human occurs distortions or blends in different human parts (e.g. the mix of the shirt/pants or the hands/pants). We also conduct experiments on the general text-to-image diffusion model GLIDE, which is hard to generate fine-grained features when the text expression is complicated. In addition, ablation experiments are also conducted. As shown in Figure 6, the models without SMR and MCA cannot represent the appearance style (e.g. denim) correctly and align the image features with the text descriptions suitably (e.g. the shirt partly appearing denim and the pants partly appearing black). In
summary, HumanDiffusion can use open-vocabulary descriptions with semantic poses to flexibly generate high-quality human images in free-style.

4.1.3 User Study

As Figure 8 shows, the User Study evaluation is performed to evaluate the quality of the generated images. 50 users are presented with 20 groups of results, in which each group has three images generated by Text2Human GLIDE and HumanDiffusion. They asked to rank 1) image photorealism and 2) caption similarity of the generated images (The scores rank is 3/4/5, higher is better). As a result, HumanDiffusion obtains higher scores for both the photorealism and captions similarity.

4.2. Semantic-Driven Person Image Generation

4.2.1 Semantic-Driven Quantitative Results

As shown in Table 2, HumanDiffusion achieves satisfactory scores in the three indicators. HumanDiffusion outperforms Pix2Pix [Wang et al., 2018], SPADE [Park and Liu, 2019] and LatentDiffusion (LDM) [Rombach et al., 2022], and exceeds the Text2Human with an improvement of on LPIPS and MSSSIM. The lightweight of our model limits the quality of generation to some extent, which greatly simplifies the training procedure.

Table 2. Quantitative comparisons on Semantic-Driven Person Image Generation with DeepFashion dataset.

| Methods       | FID↓ | LPIPS↓ | MS-SSIM↑ |
|---------------|------|--------|----------|
| Pix2Pix       | 38.46| 0.2164 | 0.7293   |
| SPADE         | 40.53| 0.2327 | 0.7668   |
| Latent Diffusion | 36.58| 0.2143 | 0.7436   |
| Text2Human    | 30.89| 0.2045 | 0.7414   |
| HumanDiffusion| 31.28| 0.2001 | 0.7736   |

4.2.2 Semantic-Driven Qualitative Results

As Figure 7 shows, Pix2Pix, SPADE, and LDM can synthesize person images in the target pose, but in low quality with distortion in details. Text2Human and HumanDiffusion generate images in high quality, while HumanDiffusion can generate diverse images (e.g. a person with several similar but different appearances in the same pose). HumanDiffusion also performs better in details like the edge of clothes, and the distinction between clothing parts and body parts, which proves that HumanDiffusion performs satisfactorily in image quality and structural rationality.

4.3. Human Partially Editing

In this section, we will introduce the performance of our method in editable applications. Different from previous generation models, we propose a more flexible method that allows using text or semantic to edit human appearances. As shown in Figure 9, we can change a man in long pants to a man in shorts by editing the pants semantic map. Slight edit on a woman in rompers can change her to wear a dress in the same style and same pose. Our method can edit local areas while preserving the rest regions unchanged.

Meanwhile, we can also edit human images under text guidance. For example, a man’s jacket color is aimed to change from grey to blue. Inspired by Blended-Diffusion [Avrahami et al., 2022], we perform a global denoising process and a local denoising process simultaneously which aims to decrease the distance between the prompts (‘blue’) and the mask image (jacket) features in CLIP distance. Stitching operations are performed on them which helps to edit the mask region and preserve the rest unchanged. As shown in Figure 10, we can generate edited images conforming to the text descriptions.

4.4. Complicated Person Generation

Existing human generation methods mostly perform poorly in complicated person generation (e.g. Occlusion situations: the person showing a side face or standing with
legs crosses). As shown in Figure 11, HumanDiffusion shows better results compared to Text2Human. As shown in red dotted boxes, in the cross-leg example, Text2Human mixes the two legs. The side face generation and hands details shown in green dotted boxes also perform unnaturally distortion. In contrast, HumanDiffusion can identify and generate partial crossover areas and natural side faces.

5. Limitations

There are two main limitations of our method. First, there are blurred textures on some little partial parts (e.g. face, shoes), since our method concentrates more on body appearances and the SMR does not set specific memory blocks for all regions. Second, some images have a cartoon style, which can be treated as a kind of blur. This is probably caused by the inherent training of diffusion. Please see more details in Supplementary Materials.

6. Conclusion

In this paper, we propose a diffusion-based text-driven person image generation framework called HumanDiffusion. We design two collaborative modules, the Multi-scale Cross-modality Alignment module and the Stylized Memory Retrieval module to align and refine the image features from coarse to fine. HumanDiffusion employs a compact architecture with end-to-end training, which could use open-vocabulary descriptions to generate person images in free-style and provide a more flexible and controllable way for real-world applications.
[Li et al., 2020] Li, S., Zhao, Y., Varma, R., Salpekar, O., Noordhuis, P., Li, T., Paszke, A., Smith, J., Vaughan, B., Damania, P., et al. (2020). Pytorch distributed: Experiences on accelerating data parallel training. arXiv preprint arXiv:2006.15704.

[Li et al., 2019] Li, Y., Huang, C., and Loy, C. C. (2019). Dense intrinsic appearance flow for human pose transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3693–3702.

[Liao et al., 2022] Liao, W., Wu, L., Zheng, F., Liu, L., and Wang, M. (2022). Text-to-image generation with semantic-spatial aware gan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18187–18196.

[Liu et al., 2022] Liu, D., Wu, L., Zheng, F., Liu, L., and Wang, M. (2022). Verbal-person nets: Pose-guided multi-granularity language-to-person generation. IEEE Transactions on Neural Networks and Learning Systems, 1, 2, 3.

[Liu et al., 2021a] Liu, T., Zhang, J., Nie, X., Wei, Y., Wei, S., Zhao, Y., and Feng, J. (2021a). Spatial-aware texture transformer for high-fidelity garment transfer. IEEE Transactions on Image Processing, 30:7499–7510.

[Liu et al., 2021b] Liu, X., Park, D. H., Azadi, S., Zhang, G., Chopikyan, A., Hu, Y., Shi, H., Rohrbach, A., and Darrell, T. (2021b). More control for free! image synthesis with semantic diffusion guidance. arXiv preprint arXiv:2112.05744.

[Liu et al., 2016] 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 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[Ma et al., 2018] Ma, L., Sun, Q., Georgoulis, S., Van Gool, L., and Schiele (2018). Disentangled person image generation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 99–108.

[Meng et al., 2021] Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. (2021). Sdeep: Guided image synthesis and editing with stochastic differential equations. In International Conference on Learning Representations.

[Miao and Wei, 2020] Miao, J. and Wei, Y. (2020). Memory aggregation networks for efficient interactive video object segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10366–10375.

[Nichol et al., 2021] Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., and Mishkin, P. (2021). Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741.

[Nichol and Dhariwal, 2021] Nichol, A. Q. and Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. In International Conference on Machine Learning, pages 8162–8171. PMLR.

[Park and Liu, 2019] Park, T. and Liu, M.-Y. (2019). Semantic image synthesis with spatially-adaptive normalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2337–2346.

[Patashnik et al., 2021] 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, pages 2085–2094.

[Pumarola et al., 2018] Pumarola, A., Agudo, A., Sanfeliu, A., and Moreno-Noguer, F. (2018). Unsupervised person image synthesis in arbitrary poses. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8620–8628.

[Radford et al., 2021] 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, pages 8748–8763. PMLR.

[Ramesh et al., 2022] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125.

[Ramesh et al., 2021] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., and Chen, M. (2021). Zero-shot text-to-image generation. In International Conference on Machine Learning, pages 8821–8831. PMLR.

[Ren et al., 2022] Ren, Y., Fan, X., Li, G., Liu, S., and Li, T. H. (2022). Neural texture extraction and distribution for controllable person image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13535–13544.

[Rombach et al., 2022] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10684–10695.

[Sarkar et al., 2021] Sarkar, K., Liu, L., Golyanik, V., and Theobalt, C. (2021). Humangan: A generative model of human images. In 2021 International Conference on 3D Vision (3DV), pages 258–267. IEEE.

[Shen et al., 2018] Shen, Y., Li, H., Yi, S., Chen, D., and Wang, X. (2018). Person re-identification with deep similarity-guided graph neural network. In Proceedings of the European conference on computer vision (ECCV), pages 486–504.

[Sheynin et al., 2022] Sheynin, S., Ashual, O., Poljak, A., Singer, U., Gafni, O., Nachmani, E., and Taigman, Y. (2022). Kmdiffusion: Image generation via large-scale retrieval. arXiv preprint arXiv:2204.02849.

[Siarohin et al., 2018] Siarohin, A., Sangineto, E., and Lathuiliere, S. (2018). Deformable gans for pose-based human image generation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3408–3416.

[Song et al., 2020a] Song, J., Meng, C., and Ermon, S. (2020a). Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502.
[Song and Ermon, 2019] Song, Y. and Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. Advances in Neural Information Processing Systems, 32. 2

[Song et al., 2020b] Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2020b). Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456. 2

[Sun et al., 2022] Sun, J., Deng, Q., Li, Q., Sun, M., Ren, M., and Sun, Z. (2022). Anyface: Free-style text-to-face synthesis and manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18687–18696. 1

[Wang et al., 2018] Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., and Catanzaro, B. (2018). High-resolution image synthesis and semantic manipulation with conditional gans. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8798–8807. 8

[Wang et al., 2022a] Wang, W., Bao, J., Zhou, W., Chen, D., Chen, D., Yuan, L., and Li, H. (2022a). Self-supervised correlation mining network for person image generation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7703–7712. 2

[Wang and Simoncelli, 2003] Wang, Z. and Simoncelli, E. P. (2003). Multiscale structural similarity for image quality assessment. In The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, volume 2, pages 1398–1402. Ieee. 6

[Weston et al., 2014] Weston, J., Chopra, S., and Bordes, A. (2014). Memory networks. arXiv preprint arXiv:1410.3916. 5

[Xu et al., 2021] Xu, R., Guo, M., Wang, J., Li, X., Zhou, B., and Loy, C. C. (2021). Texture memory-augmented deep patch-based image inpainting. IEEE Transactions on Image Processing, 30:9112–9124. 2

[Xu and Chen, 2021] Xu, X. and Chen, Y.-C. (2021). Text-guided human image manipulation via image-text shared space. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1, 2

[Zhang et al., 2021a] Zhang, J., Li, K., and Lai, Y.-K. (2021a). Pise: Person image synthesis and editing with decoupled gan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7982–7990. 2

[Zhang et al., 2021b] Zhang, J., Siarohin, A., Tang, H., Chen, J., Sangineto, E., Wang, W., and Sebe, N. (2021b). Controllable person image synthesis with spatially-adaptive warped normalization. arXiv preprint arXiv:2105.14739. 3

[Zhang and Yang, 2022] Zhang, P. and Yang, L. (2022). Exploring dual-task correlation for pose guided person image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7713–7722. 2

[Zhang et al., 2018] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 586–595. 6

[Zhang and Zhao, 2021] Zhang, S. and Zhao, W. (2021). Keypoint-graph-driven learning framework for object pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1065–1073. 1, 2

[Zhou et al., 2022] Zhou, X., Yin, M., Chen, X., Sun, L., Gao, C., and Li, Q. (2022). Cross attention based style distribution for controllable person image synthesis. arXiv preprint arXiv:2208.00712. 2

[Zhou et al., 2021] Zhou, Y., Zhang, R., Chen, C., Li, C., Tensmeyer, C., Yu, T., Gu, J., Xu, J., and Sun, T. (2021). Lafite: Towards language-free training for text-to-image generation. arXiv preprint arXiv:2111.13792. 1

[Zhu and Pan, 2019] Zhu, M. and Pan, P. (2019). Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5802–5810. 3

[Zhu et al., 2019] Zhu, Z., Huang, T., Shi, B., Yu, M., Wang, B., and Bai, X. (2019). Progressive pose attention transfer for person image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2347–2356. 1, 2