UPGPT: Universal Diffusion Model for Person Image Generation, Editing and Pose Transfer

Soon Yau Cheong
University Of Surrey
s.cheong@surrey.ac.uk

Armin Mustafa
University of Surrey
armin.mustafa@surrey.ac.uk

Andrew Gilbert
University of Surrey
a.gilbert@surrey.ac.uk

Abstract

Text-to-image models (T2I) such as StableDiffusion have been used to generate high quality images of people. However, due to the random nature of the generation process, the person has a different appearance e.g. pose, face, and clothing, despite using the same text prompt. The appearance inconsistency makes T2I unsuitable for pose transfer. We address this by proposing a multimodal diffusion model that accepts text, pose, and visual prompting. Our model is the first unified method to perform all person image tasks - generation, pose transfer, and mask-less edit. We also pioneer using small dimensional 3D body model parameters directly to demonstrate new capability - simultaneous pose and camera view interpolation while maintaining the person's appearance.

1. Introduction

Generating humans from text and/or pose is a challenging problem in computer vision. The methods can be classified into two categories, (1) image generation (synthesis) and (2) pose transfer and editing. Image generation can be unconditional or conditioned on other information *e.g.*, pose and text. Pose-guided image generation, conditions on pose (keypoints, skeleton image, heatmap, body mesh) to generate images [14, 7, 25, 1]; and text-to-image models such as DALL-E[28, 27] and [43, 40, 51, 37, 42, 34]. Pose or text to image is a one-to-many mapping. It can create a person with vastly different appearances even given the same conditions - *e.g.*, a person in the same pose but wearing other clothing or shades of color from the word "red shirt." The ambiguity and inconsistency prohibit them from being used to perform image editing, which requires the maintenance of the visual appearance of all other aspects of the image apart from the elements or regions to be edited. Some newer methods [5, 15, 46] use pose and text to exert further control. However, the effect is still limited by the inherent ambiguity of

Code and demo available at github.com/soon-yau/upgpt

Figure 1: UPGPT can perform all person image generative tasks: (a) text and pose guided image generation, (b) fine-grained, mask-less region editing with text, (c) style and appearance transfer, (d) pose transfer followed by edit.

these modalities and is hence unsuitable for image editing.

The other category is image editing, for tasks such as changing the clothing, human pose, or face. Most pose-guided image generation literature fall into this category, performing pose transfer to transfer a person's appearance from a source image to the pose of a target image. However, we prefer the term *edit* to encompass other forms of modification, including using text or modifying the pose parameters directly, rather than having to *transfer* them from the other image. Pose transfer models [22, 35, 41] use both human pose and a source image as conditions for the generative image model where visual information of a source

image serves as a stable condition to encourage the models to maintain a person's appearance in the generated image. [52, 44, 29, 50] have extended capabilities that could also transfer texture, clothing shape, or both, *i.e.*, appearance transfer, but no single model can perform all those tasks. More importantly, they all need to train on source-target image pairs and can not generate a new image without a source image. To bridge the gap between the two categories, we propose *UPGPT* to perform both generation and edit tasks using a single trained universal model, and image sampling pipeline, as seen in Figure 1.

In our research, we discovered four underlying problems in person image generation and editing that have yet to be addressed: **(1)** Existing methods cannot interpolate human pose due to the inherent limitation of the chosen pose representations. 2D body segmentation map (parsing map) and body mesh are dense representations (pixel and voxel) and cannot be interpolated. To interpolate 2D keypoint points and their derivatives (skeleton image, heatmap), they must first be mapped into 3D space, which is a difficult task on its own, before performing the interpolation in 3D space, then project back into 2D keypoints. We break away from the tradition by using pose parameters of SMPL[19], a 3D body model that represents pose by rotation of body joints. Then, performing linear interpolation on the SMPL parameters produces pose interpolation using our model. **(2)** Existing person image editing methods require parsing maps, which is difficult for users to create or edit by hand. Furthermore, their methods are typically constrained to transferring information from a single modality. To address this challenge, our method allows text or drag-and-drop of the reference image or a combination of them to perform convenient and fast image editing. **(3)** Missing information from the source image. For example, when a target image expects a full-body person, the source image only contains a partial view where the lower part is not visible, as shown in Figure 2. This leaves a question of whether the model should generate short pants, long pants, a dress, a sneaker, high heels, or leather shoes. **(4)** A person's appearance can change in the target image *e.g.*, a person wearing a jacket in the source image may have it taken off in the target. Existing methods rely solely on the source image to provide all the information. Still, they can fail to generate desired or correct results if the information is incomplete or wrong, as shown in Figure 6. We address problems 3 and 4 by adding a new modality - text to enrich the information source and to reduce and correct errors. The text description of the expected outcome can work as a way to filter out unwanted information (not wearing a jacket) or to fill in missing information (to generate pants or skirts).

Table 1 compares the capabilities of the two main person image generation methods, and our proposed method combines all the key features. In summary, the main contri-

Figure 2: Pose transfer is an ill-posed problem: Often, the source image does not contain all information for the target pose *i.e.*, in this figure, the pant. Compared to existing methods (PISE[44], ADGAN[23], DPTN[48], NTED[29], CASD[50]), our method can create the desired result by utilizing additional multimodal information.

butions of our papers are:

1. A unified framework that can simultaneously perform person image generation, editing, and pose transfer tasks.

2. The provision of zero-shot, mask-less image generation and editing with text.

3. The use of 3D parametric body model parameters to demonstrate the first simultaneous pose and camera view interpolation.

	Pose Transfer	Text-Pose-to Person Image	UPGPT (Ours)
Pose Edit	✓	✗	✓
Appearance Edit	✓	✗	✓
Texture Edit	✓	✗	✓
Create from Text	✗	✓	✓
Edit with Text	✗	✓	✓
Pose Interpolation	✗	✗	✓

Table 1: Comparing the superset capabilities of pose transfer[52, 41, 30, 23, 44, 50, 29], text-pose-to-image [5, 46, 15] and our method. Our unified method can perform all the person generation and edit tasks and introduce a new capability of pose interpolation.

2. Related Works

Diffusion Models (DM) [12, 6] have shown superior image quality and text-guided capability. In training, the DM gradually adds noise to the image until it becomes random noise; this process is known as forward diffusion. The diffused random noise act as latent variables and is denoised progressively to generate an image in image sampling; this progress is known as reverse diffusion. Typically, a UNet[32] is used to learn to produce the denoising signal. Most methods[27, 34, 24] use the classified-free approaches[17] to find the direction between the conditional and unconditional in the latent space, which is to be applied in sampling time to guide the model towards the conditioning direction. However, denoising every image pixel can be

computationally expensive; therefore, LDM[31] proposed a two-stage process. It first trained a variational autoencoder (VAE) [18] to encode the image into smaller dimensional latent variables, and the DM learned to produce the VAE latent variables. DMs could provide image editing by performing text-guided diffusion on regions defined by segmentation mask [24, 27, 34, 3, 2]. Dreambooth[33] shows that they could encode a person's face into a text token and use a DM to generate the person in a different scene. More recently, [10] proposed a mask-less edit of coarse objects by learning the region from the attention map. Our method achieves mask-less editing by learning and disentangling a person's appearance.

Pose Guided Image Generation. Ma *et al*. [22] was among the first literature on pose transfer; they concatenated source images with the target pose heatmap and used them as input conditions to a GAN[8]. Starting from PATN[52], models take pose from both the source and image. Yang *et al*. [41] detected and cropped out the person's face and used that as an additional image condition within the network for a more fine-grained detailed generation. In addition to human pose, ADGAN [23] uses a human parsing map to segment the body parts of the source image to extract their style codes. This allowed them to change the style or texture of clothing region. However, as the shape of the person and clothing is bounded by the segmentation map of the source image, the image edit is limited to only texture transfer. To overcome this issue, PISE [44] and SPGNET [21] trained a separate network to generate a parsing map of the target pose, which they edited before feeding into the image generator. Allowing the changing of clothing shape *e.g*., from short sleeve to long sleeve, but they cannot perform texture transfer simultaneously. While NTED[29] and CASD[50] demonstrated transfer of the entire clothing pieces, they do not provide a method to transfer only the texture. DPTN[48] uses two paths - source-to-source and source-to-target, while we require only one path for both trainings. Unlike our approach, existing methods can only edit a subset of clothing texture, shape, or appearance (texture and body), but not all of them. [1] uses SMPL body mesh, which is more computationally expensive to process than our method, which uses only 72 parameters. Concurrent to our work, PIDM[4] shows clothing style interpolation by interpolating the DM's noises, but they could not perform pose interpolation.

Text-Guided Image Generation. There exist text-to-image models since the early days of GANs [43, 40, 51], to transformer[38]-based DALL-E[28] and diffusion models[24, 27, 31, 34]. However, they do not provide precise control over human pose or fine-grained appearances. KPE[5] created the first text-and-pose-guided image generative model that encodes body keypoints into transformer tokens as conditions. Although it can generate ac-

curate poses, as text is a weak condition, it cannot provide fine-grained appearance control and consistency for pose transfer. Using a parsing map and hierarchical autoencoder to encode different body regions, Text2Human[15] offered more fine-grained appearance control. Still, it could not specify person and clothing attributes not labeled in the text description, notably the clothing color. HumanDiffusion[46] segments and encodes each clothing item with CLIP image encoder into style code and uses a fixed-size database to store the embedding of fashion styles. During sampling, they use either CLIP image or text embedding, but not both, to retrieve the closest embedding from the database. Although this allows them to control the clothing color using text, their method entangles the clothing type, color, and texture pattern into a finite number of combinations. In contrast, our method offers disentanglement and allows users to edit each clothing attribute independently using combination of image and text. The existing generative methods could not consistently generate images for pose transfer or appearance editing. More recently, ControlNet [47] adds pose guidance to DM, but it cannot ensure appearance consistency due to the lack of visual conditioning.

3. Methodology

The primary motivation of our proposed method is to fully disentangle a person's image into content and style represented by pose, text, and image features. We can independently edit and mix the different modalities at source to provide fine-grained person image generation and editing. Figure 3 illustrates the overall architecture of UPGPT. The first step is to extract the person's information from images and text in the form of features and encode them into conditioning embeddings. The second step is to fuse the embeddings within Multimodal Fusion Block (MFB) to provide conditioning to the UNet of the DM. The figure shows the training pipeline for the pose transfer task with the source-target image pair at the input. However, this can be repurposed for image generation tasks using the same image as the source and target image. Existing person image generation methods [5, 15, 46] use only individual images in training, while image pairs are necessary for pose transfer methods [52, 41, 30, 23, 44, 50, 29, 4]. Our novel architecture allows us to use individual and paired images to increase the training sample size. The following section describes the proposed method in detail.

3.1. Multimodal Feature Representation

Our model uses three modalities: pose, image, and text. We further divide the text into context text and style text. Overall, a person's image is disentangled into content represented by pose and context text; and style as defined by style text and image.

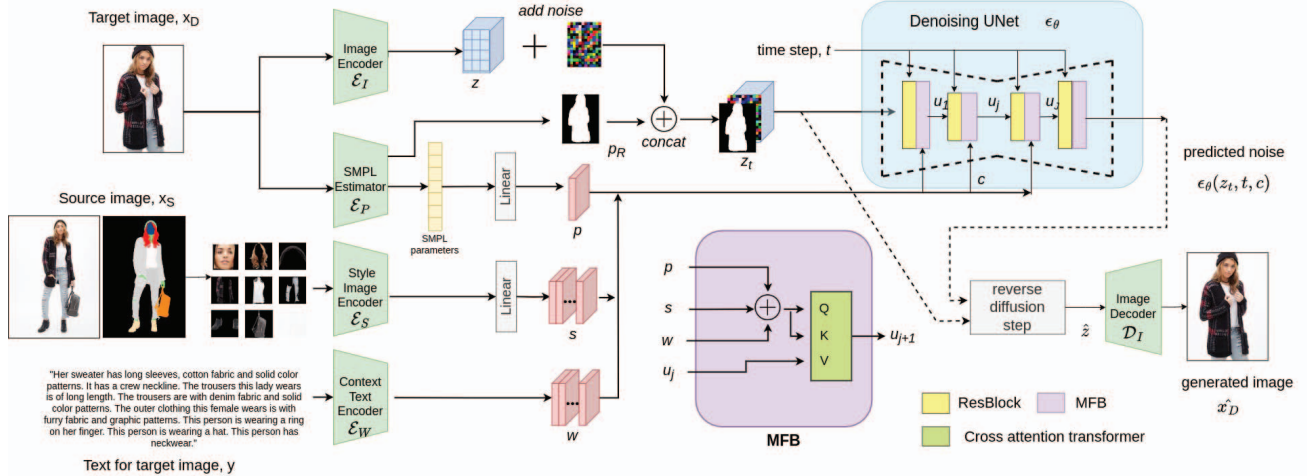**Image Latent.** We encode the target image $x_D \in$

Figure 3: Overview of our proposed UPGPT architecture. In training, we encode pose, style image, and context text into embeddings that go to the Multimodal Fusing Block (MFB) for fusing. The output of MFB is used as a condition in UNet to predict the noise needed to denoise the image's latent. In sampling, the image encoder decodes the denoised latent $\hat{z}$ into pixel space.

$\mathbb{R}^{H \times W \times 3}$ using VAE's [18] encoder into the latent variables $\mathcal{E}_\mathcal{I}(x_D) = z \in \mathbb{R}^{\frac{H}{f}, \frac{W}{f}, d_V}$ where d_V is the VAE's channel dimension, and f is a downsampling factor in the power of two, and x_D and z are only needed in training. In image sampling, the trained DM generates a new image latent $\hat{z}$, to be decoded by the VAE decoder into pixel space $\hat{x_D} = \mathcal{D}_\mathcal{I}(\hat{z})$. Smaller f e.g., 4 gives higher spatial resolution but quadruples the latent size from $f = 8$ and thus increases computational effort considerably. Although large f is more computationally frugal, the resulting z has a smaller spatial dimension, which will store more visual details for the same pixel patch. As a result, a small face in a full-body image can appear blurry after image reconstruction $\mathbb{D}_I(\mathcal{E}_\mathcal{I}(x_D))$.

SMPL Pose. We use [45] as pose estimator $\mathcal{E}_\mathcal{P}$ to create an embedding based on the SMPL parameters from the target image x_D. The 72 SMPL parameters represent three axis-angle rotations of 24 body joints, ten body shape parameters, and three camera parameters. The camera view of an image is determined by the body's vertical axis rotation parameter and the camera parameters. Each of the three camera parameters in Cartesian coordinate axes determines horizontal translation, vertical translation, and zooming. The SMPL parameters are flattened and projected with a linear layer to $p \in \mathbb{R}^{1 \times d}$ where d is the context text embedding channel dimension. Experiments show that the SMPL's camera parameters are insufficient to ensure the person's correct horizontal position. Therefore, we concatenate a silhouette mask $p_R \in \mathbb{R}^{\frac{H}{f}, \frac{W}{f}}$ at the UNet input to reinforce the pose conditioning, we call it as **reinforced person mask (RPM)**. RPM only needs to be a coarse mask; this differs from [23, 44, 21, 50], which requires a detailed body part segmentation map. We used binary silhouette mask in our main experiments but tried other methods as discussed

further in Section 4.5.

Style Image. From a source image x_S, we use a segmentation map to segment the person into 9 fine-grained semantic regions i.e., head, hair, headwear, background, top, bottom, outwear, shoes, and bag. Each of the segmented regions is cropped and resized. We call this style image, and we use it as a condition for the person's appearance style. Unlike conventional methods that perform segmentation in run time, we do it in the data preparation stage and store the style regions. This provides image editing flexibility by simply changing the style image files. We do not use source images anymore after obtaining the style images. We treat a person's identity as one of the styles determined by face and hairstyle images. We use a separate face detector to normalize the face - align the face to an upright position. If an occluded face is not detected, we replace it with another normalized face image from the same person if it is available. We encode the style images with a pre-trained CLIP [26] image encoder $\mathcal{E}_\mathcal{S}$ before projecting it with a linear layer into $s \in \mathbb{R}^{N \times d}$ where N is the number of style regions defined for a person.

Style Text. CLIP [26] trains an image encoder and text encoder jointly on image-text pairs, with a common embedding for both modes aiming to be close to each other in the CLIP embedding space. For example, the CLIP embedding of the text "a red shirt" and an image of a red shirt should be close in terms of Euclidean distance. We use this to create a zero-shot learning method through editing with text. Like us, HumanDiffusion[46] uses CLIP image encoding in training, but they can only use either text or image to control image sampling, while we can use either or both modalities. Also, we use two different text conditions - content and style to provide better disentanglement and finer control. Figure 4 shows how we can mix the style images and

texts in image sampling. Style text provides a fast and convenient way to control the clothing texture and color, while we can use style images to dictate specific appearances such as face and color shade.
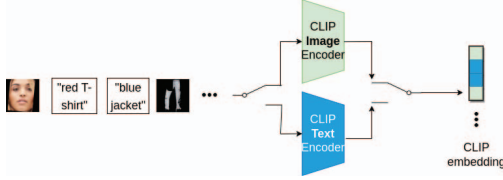

Figure 4: We can mix-and-match a combination of image (green) and text (blue) embedding in sampling time.

Content Text. The content text describes the content of the target image *e.g.*, gender, clothing shapes, and fabrics. We use a pre-trained LLM (large language model) transformer [13] for text encoding. We take the transformer's last layer feature as our content text embedding $\mathcal{E}_{\mathcal{W}}(y) = w \in \mathbb{R}^{l \times d}$ where l is the maximum text token length.

3.2. Conditional Diffusion Model

The DM training process consists of a sequence of time steps $t = 1...T$, where Gaussian noise ϵ is scaled using a noise schedule [12] and added to an image latent variable z to produce a noisy version. This concatenates with p_R to produce z_t, fed into the input of a denoising UNet ϵ_θ. We propose to condition using our MFB block to concatenate $\oplus$ pose p, text w and style embedding s and perform cross-attention with UNet's ResBlock output at every level u_j where j is layer number.

$$c = p \oplus s \oplus w, Q = \phi_Q(c), K = \phi_k(c), V = \phi_V(u_j) \quad (1)$$

where ϕ performs 1×1 convolution layers for projection into u_j's channel dimension d_j and flatten to 1-dimension.

$$CrossAtten(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_j}})V \quad (2)$$

We train the UNet by using MSE loss on predicted noise $\epsilon_\theta(z_t, t, c)$:

$$\mathcal{L}_{\mathcal{MSE}} := \mathbb{E}_{z, p_R, c, t, \epsilon \sim \mathcal{N}(0,1)} \left[\|\mathcal{W} \odot (\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2) \right] \quad (3)$$

Where $\odot$ is element-wise multiplication and $\mathcal{W} \in \mathbb{R}^{\frac{H}{f}, \frac{W}{f}}$ is loss weight we add to the standard diffusion loss. In addition to the primary loss, many GANs[44, 30, 50, 29] use perceptual loss[16], which extract features from image pixels. However, a single training step in the DM does not generate an image; therefore, we cannot directly use additional losses that require image pixels. Consequently, we use a loss weight $\mathcal{W}$, a 2D tensor with the same dimension as the image latent, to assign different weights to the loss. This helps to regulate the training under challenging regions such as face and hands.

Task\Condition	Styles	Content Text	Pose
Generate	source	source	source
Texture Edit	style image/	source	source
Shape Edit	source	target/edit	source
Appearance Edit	style image/	target/edit	source
Pose Transfer	source	target	target

Table 2: Starting from image generation using information from the source image, the table shows how our method can perform various tasks using different conditioning combinations.

3.3. Generation, Transfer & Editing of Images

Unlike previous pose transfer work, we do not need to use a segmentation map or any reference person image when sampling a new image. We create a new random image latent z_0 to begin the sampling process. Progressively in each time step t, the image latent is denoised using the reverse diffusion step as described by [12] to produce a less noisy image latent $\hat{z}_t = G(z_t, t, c)$. After the T steps, the denoised $\hat{z}$ is decoded by the VAE decoder $\mathcal{D}_{\mathcal{I}}(\hat{z})$ to create an image in pixel space. We use the same pipeline for all the tasks by changing only the conditioning.

To adjust the clothing texture and color, we can either do a texture transfer by using a style image or by replacing the style embedding for that clothing with style text, all without a segmentation mask. Due to the suitable disentanglement property of our method, this changes only the texture and color but not the clothing shape, as demonstrated in the left image in Figure 1(c). If we fix the style condition and change only the context text *e.g.*, from "long sleeve" to "short sleeve," it will only change the sleeve length while maintaining the clothing texture. We can modify the content text and style for appearance edit/transfer, which change/copies both the shape and styles. To perform pose transfer, we replace the pose of the source image with one from the target image; this would produce results similar to existing pose transfer methods. On top of that, we use context text from the target image that better describes the desired appearance to generate images with clothing appearance more faithful to the target image. The different configurations are summarized in Table 2, and some image examples are shown in Figure 1.

4. Experiments

We performed experiments on two tasks: (1) text-pose guided image generation and (2) pose transfer. Both use the same model architecture but different image resolutions and subsets of the DeepFashion dataset [53].

Implementation Details. We train our model using AdamW optimizer [20] at a learning rate of 5×10^{-5}, batch size of 24, and loss weight, W (Equation 3) used is face=8.0, arms=2.0, background=0.5 and 1.0 for others,
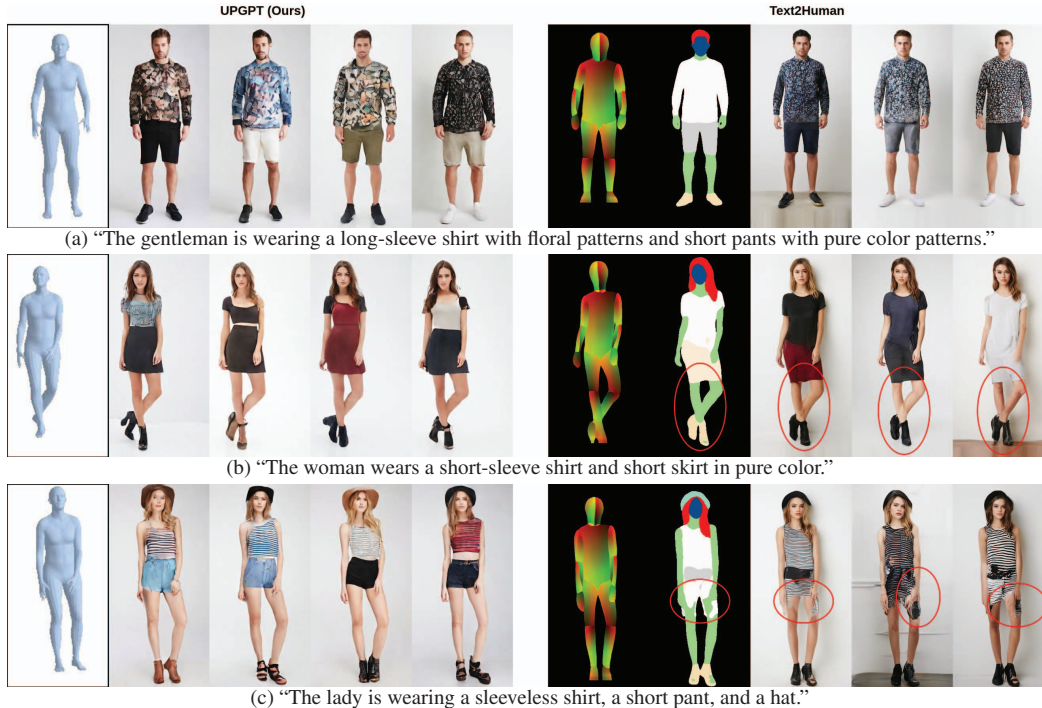
(a) "The gentleman is wearing a long-sleeve shirt with floral patterns and short pants with pure color patterns."

(b) "The woman wears a short-sleeve shirt and short skirt in pure color."

(c) "The lady is wearing a sleeveless shirt, a short pant, and a hat."

Figure 5: *(Zoom in to view full* 512×256*)* resolution. (a) We generate a variety of clothing types and texture patterns directly from SMPL pose parameters while Text2Human has additional stage to create parsing map from pose (DensePose[9]). (b) Text2Human tend to generate blended crossed legs when the parsing map overlapped. (c) Using vocabulary outside of Text2Human limited dictionary can result in defective parsing map and hence erroneous final image.

and a silhouette mask is used for reinforced pose mask. Our model is trained with $T = 1000$ noising steps and a linear noise schedule.

Evaluation Metrics. We use *LPIPS*[49] and *SSIM* [39] to measure the similarity between the generated image and target image in the pose transfer task. LPIPS uses pre-trained VGG[36] to calculate the perceptual similarity, while SSIM measures the similarity by considering the images' luminance, contrast, and structure. For the text-pose guided image generation task, clothing color changes can significantly impact the similarity score, even if it looks realistic. Therefore, instead of comparing individual images, we use Frechét Inception Distance (*FID*)[11] to measure the distribution of two groups - ground truth and generated images.

4.1. Text-Pose Guided Image Generation

Method	FID↓
†HumanDiffusion[46]	30.42
Text2Human[15]	24.52
UPGPT(Ours)	**23.46**

Table 3: Quantitative result on DeepFashion Multimodal dataset on text-and-pose guided image generation. † taken from [46].

We use the DeepFashion Multimodal dataset proposed by Text2Human [15] in which a segmentation map and text description accompany each image. We train on the resolu-

tion 512×352. We follow Text2Human's data split and crop the generated images into 512×256. The baseline methods [46, 15] cannot control clothing color, which would hugely affect evaluation scores. For a fair comparison, we train our models without clothing style image embedding.

Table 3 shows our method achieving the best FID score against the baselines. Next, we perform some qualitative analysis. HumanDiffusion[46] does not provide code to reproduce their results, but their paper shows blurry images with color saturation. Both us and Text2Human can generate high quality images, as shown in Figure 5a and Figure B.1 in the appendix, but there are a few shortcomings with the latter. Text2Human cannot generate images directly from the pose, and it must first generate a parsing map from the pose and text. As also observed by [46], we found that they systematically exhibit blended crossed leg when parsing map overlapped (Figure 5b). Parsing maps can also induce gender bias, as detailed in the appendix. Also, Text2Human has limited text capability. Their model was trained on categorical labels and added text-to-category mapping later. Therefore, vocabulary falling outside of their dictionary can generate the wrong parsing map. This is demonstrated in Figure 5c. The word *pant* rather than *pant**s*** was used in the text prompt, and that causes the skin (green) and top clothing (white) to smear into bottom clothing (gray). The following sections will show our superior visual and text prompting capability.

Figure 6: *(Zoom in to view)* Pose transfer from (1) source image into the (9) pose target in which the jacket is removed. Reference methods PISE[44], ADGAN[23], DPTN[48], NTED[29], CASD[50] blend the top wear and jacket to generate the wrong clothing (2-6), while ours (7) create clear separated jacket from top wear, matching the source image appearance. Conditioning on the content text that correctly describes the target image, we create the final pose transfer result in (8) matching the ground truth (9) appearances. (10) and (11) show we can perform consecutive texture and appearance transfers with texts. In (12), we show how to perform texture and identity transfer using style images while still conditioning on the previous style text edit.

4.2. Pose Transfer

We use DeepFashion[53] In-shop Clothes Retrieval dataset for the pose transfer task. Using the given train-test split of individual images (48675 and 4039, respectively), PATN[52] proposes a pose transfer dataset of about 102k image pairs for training and 8570 pairs for testing. Given our model architecture's flexibility to support individual and image pairs in training, we combine both as our training dataset. As the Inshop subset does not provide a text description of images, we use the text labels from the Multimodal subset, which cover most of the samples in Inshop. We resize the image to 256×176, maintaining the same aspect ratio. We also combined the fine-grained segmentation map from both subsets. However, a small number, about 5% of Inshop test image pairs, either have incomplete text or segmentation maps or do not contain humans; we excluded these from the test set. We evaluate our and reference methods using the same reduced test set to obtain the fair quantitative results in Table 4.

Although our method is not designed explicitly for the pose transfer task alone, we near state-of-the-art results;

Method	FID↓	LPIPS↓	SSIM↑
ADGAN[23]	20.025	0.2289	0.6856
PISE[44]	17.799	0.2273	0.6781
DPTN[48]	16.686	0.2192	0.6958
CASD[50]	10.439	**0.1777**	**0.7131**
UPGPT(ours)	9.427	0.1886	0.6970
NTED[29]	**8.813**	0.1814	0.7011
†UPGPT(ours)	7.876	0.1766	0.7276

Table 4: Quantitative results on pose transfer task. † compare the generated images against images reconstructed by VAE.

we found that small faces in our generated images can appear blurry due to the inadequacy of VAE in capturing rich details in small faces. In other words, an image x reconstructed $\mathcal{D}_\mathcal{I}(\mathcal{E}_\mathcal{I}(x))$ by VAE can have a blurry face even if our model produces a perfect image latent. To confirm this, we compare our generated images against the images reconstructed from the ground truth images rather than the ground truth images, and the scores improve significantly to top the performance table.

Apart from that, our method produces realistically look-

ing images and excels in utilizing all modalities when information in the source image is incomplete or incorrect. This is best demonstrated in the pose transfer task in Figure 6, where the jacket in the source image (1) is removed from the target image (9). Even assuming the person still has their jacket on, existing methods (2-6) often fail to distinguish between the jacket and the top wear, blending the style and texture to create incorrect clothing. In contrast, our results (7) show clear distinguishment, resembling the source image appearance. UPGPT blocks out the jacket in the image generation process by conditioning on the context text of the target image, creating our final pose transfer result (8) that resembles the ground truth target image (9).

4.3. Flexible Image Editing

Columns (10-12) in Figure 6 demonstrate the flexibility of our fine-grained control method. From (7), we replace the jacket style image with the style text "jacket in orange leopard pattern" to perform texture transfer (10). Our style text has good zero-shot capability, and we can use words like zebra, pandas, and oceanic instead of color. Then, we change the context and style texts in (11) to replace bottom wear with a short green skirt, changing the texture and clothing type *i.e.* appearance edit. Please note that the jacket from (10) remains in (11), showing that our approach allows for consecutive editing. This is a significant improvement from existing methods [50, 29, 4] that have demonstrated only to transfer appearance from a single image reference. In contrast, we can mix different modalities from different sources to perform flexible and fine-grained control across clothing type, texture, or both. Although it is convenient to use text to change clothing types and colors, some things are difficult to describe in words *e.g.* specific clothing patterns or the face of a particular person. Therefore, our methods also support using styles images for editing and identity transfer, as shown in (12). The pipeline of going from (1) to (12) demonstrates we can mix and match different modalities - pose, style, content text, and style images to achieve excellent fine mix-and-match in generating and editing person images.

4.4. Pose and Camera View Interpolation

We demonstrate our approach's superior pose capability and disentanglement with simultaneous pose and camera view interpolation as shown in Figure 7. The pose interpolation can be performed by linear interpolating SMPL parameters between two poses. To our best knowledge, this is the first demonstration of pose interpolation within the human image generation literature.

4.5. Ablation Study

We performed experiments to explore the importance of the reinforced pose masks (RPM) and evaluated their performances as shown in Table 5. Qualitatively, without any form of RPM, the person in generated images looks visually

Figure 7: Complex hand and camera movement achieved using linear pose interpolation.

similar to other masks apart from the occasional horizontal offset. We explore two methods: a bounding box and a mask derived from the SMPL render. Including a bounding box as a mask improves the scores compared to not having one. A further approach is to use the silhouette mask created from SMPL rendering as the segmentation input. However, the derived mask is less accurate, and the result is slightly worse than using a silhouette mask estimated from 2D images.

Reinforced Pose Mask (RPM)	FID↓	LPIPS↓	SSIM↑
w/o RPM	10.176	0.2670	0.6146
w bounding box	10.100	0.2447	0.6254
w SMPL rendering	**9.245**	0.2149	0.6604
w silhouette mask	9.427	**0.1886**	**0.6970**

Table 5: Quantitative results of ablation on reinforced pose mask.

5. Limitations

Apart from the blurry small faces discussed in Section 4.2, one of our method's limitations is that clothing textures sometimes do not match the style image *e.g.*, the stripes can have different thicknesses. This is due to the limitation of the CLIP image encoder, which does not necessarily capture fine-grained spatial detail but focuses on the overall color response.

6. Conclusion

In this paper, we proposed UPGPT, the first universal method to perform unified person image generation, editing, and pose transfer tasks. Unlike existing methods that require masks for editing, our mask-less approach provides a convenient way of fine-grained person image editing using a combination of modalities. We achieved competitive pose transfer results in comparison to the state-of-the-art methods. Also, we overcame the inadequacy of SMPL pose estimation to incorporate it into our model to improve pose disentanglement and demonstrate the first simultaneous pose and camera view interpolation in pose-guided image generation literature.

References

[1] Badour AlBahar, Jingwan Lu, Jimei Yang, Zhixin Shu, Eli Shechtman, and Jia-Bin Huang. Pose with style: Detail-preserving pose-guided image synthesis with conditional stylegan. *SIGGRAPH Asia*, 9 2021. 1, 3

[2] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Transaction on Graphics*, 6 2022. 3

[3] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 11 2022. 3

[4] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. *IEEE Conference of Computer Vision and Pattern Recognition (CVPR)*, 11 2023. 3, 8

[5] Soon Yau Cheong, Armin Mustafa, and Andrew Gilbert. Kpe: Keypoint pose encoding for transformer-based image generation. *British Machine Vision Conference (BMVC)*, 3 2022. 1, 2, 3

[6] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 2

[7] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[8] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Conference on Neural Information Processing Systems (NeurIPS)*, 2014. 3

[9] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. *IEEE Conference of Computer Vision and Pattern Recognition (CVPR)*, 2 2018. 6

[10] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *Arvix Preprint 2208.01626*, 8 2022. 3

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017. 6

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 2, 5

[13] HuggingFace. openai/clip-vit-large-patch14. 5

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[15] Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: Text-driven controllable human image generation. *SIGGRAPH*, 2022. 1, 2, 3, 6

[16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. *European Conference on Computer Vision (ECCV)*, 3 2016. 5

[17] Tim Ho Jonathan Salimans. Classifier-free diffusion guidance. *NeurIPS 2021 Workshop DGMs Applications*, 2021. 2

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *Arxix Preprint Auto-Encoding Variational Bayes*, 12 2013. 3, 4

[19] Matthew Lopper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34, 2015. 2

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR)*, 11 2017. 5

[21] Zhengyao Lv, Xiaoming Li, Xin Li, Fu Li, Tianwei Lin, Dongliang He, and Wangmeng Zuo. Learning semantic person image generation by region-adaptive normalization. *IEEE Conference of Computer Vision and Pattern Recognition (CVPR)*, 4 2021. 3, 4

[22] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. *Conference on Neural Information Processing Systems (NeurIPS)*, 2017. 1, 3

[23] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. *IEEE Conference of Computer Vision and Pattern Recognition (CVPR)*, 3 2020. 2, 3, 4, 7

[24] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *Proceedings of Machine Learning Research*, 2021. 2, 3

[25] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning (ICML)*, 2 2021. 4

[27] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *Arxiv Preprint: 2204.06125*, 4 2022. 1, 2, 3

[28] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *International Conference on Machine Learning (ICML)*, 2021. 1, 3

[29] Yurui Ren, Xiaoqing Fan, Ge Li, Shan Liu, and Thomas H. Li. Neural texture extraction and distribution for controllable person image synthesis. *IEEE Conference of Computer Vision and Pattern Recognition (CVPR)*, 4 2022. 2, 3, 5, 7, 8

[30] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H. Li, and Ge Li. Deep image spatial transformation for person image generation. *IEEE Conference of Computer Vision and Pattern Recognition (CVPR)*, 3 2020. 2, 3, 5

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 12 2022. 3

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI)*, 5 2015. 2

[33] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *Arxiv preprint 2208.12242*, 8 2022. 3

[34] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *Arxiv preprint: 2205.11487*, 5 2022. 1, 2, 3

[35] Aliaksandr Siarohin, Enver Sangineto, Stephane Lathuiliere, and Nicu Sebe. Deformable gans for pose-based human image generation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*, 9 2014. 6

[37] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. Df-gan: A simple and effective baseline for text-to-image synthesis. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8 2020. 1

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Conference on Neural Information Processing Systems (NeurIPS)*, 2017. 3

[39] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 6

[40] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 3

[41] Lingbo Yang, Pan Wang, Chang Liu, Zhanning Gao, Peiran Ren, Xinfeng Zhang, Shanshe Wang, Siwei Ma, Xiansheng Hua, and Wen Gao. Towards fine-grained human pose transfer with detail replenishing network. *IEEE Transactions on Image Processing*, 2020. 1, 2, 3

[42] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[43] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *International Conference on Computer Vision (ICCV)*, 2017. 1, 3

[44] Jinsong Zhang, Kun Li, Yu-Kun Lai, and Jingyu Yang. Pise: Person image synthesis and editing with decoupled gan. *IEEE Conference of Computer Vision and Pattern Recognition (CVPR)*, 3 2021. 2, 3, 4, 5, 7

[45] Jason Y. Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. *European Conference on Computer Vision (ECCV)*, 7 2020. 4

[46] Kaiduo Zhang, Muyi Sun, Jianxin Sun, Binghao Zhao, Kunbo Zhang, Zhenan Sun, and Tieniu Tan. Humandiffusion: a coarse-to-fine alignment diffusion framework for controllable text-driven person image generation. *Arxiv Preprint 2211.06235*, 11 2022. 1, 2, 3, 4, 6

[47] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv:2302.05543*, 2 2023. 3

[48] Pengze Zhang, Lingxiao Yang, Jianhuang Lai, and Xiaohua Xie. Exploring dual-task correlation for pose guided person image generation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3 2022. 2, 3, 7

[49] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *IEEE Conference of Computer Vision and Pattern Recognition (CVPR)*, 1 2018. 6

[50] Xinyue Zhou, Mingyu Yin, Xinyuan Chen, Li Sun, Changxin Gao, and Qingli Li. Cross attention based style distribution for controllable person image synthesis. *European Conference on Computer Vision (ECCV) IEEE Conference of Computer Vision and Pattern Rec*, 8 2022. 2, 3, 4, 5, 7, 8

[51] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4 2019. 1, 3

[52] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. *IEEE Conference of Computer Vision and Pattern Recognition (CVPR)*, 4 2019. 2, 3, 7

[53] Ziwei, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Liu Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5, 7