StyleHumanCLIP: Text-guided Garment Manipulation for StyleGAN-Human

Takato Yoshikawa[1][a], Yuki Endo[1][b] and Yoshihiro Kanamori[1][c]
[1] *University of Tsukuba, Japan*
tenten0727@icloud.com, {endo,kanamori}@cs.tsukuba.ac.jp

Keywords: StyleGAN, Text-guided Garment Manipulation, Full-body Human Image

Abstract: This paper tackles text-guided control of StyleGAN for editing garments in full-body human images. Existing StyleGAN-based methods suffer from handling the rich diversity of garments and body shapes and poses. We propose a framework for text-guided full-body human image synthesis via an attention-based latent code mapper, which enables more disentangled control of StyleGAN than existing mappers. Our latent code mapper adopts an attention mechanism that adaptively manipulates individual latent codes on different StyleGAN layers under text guidance. In addition, we introduce feature-space masking at inference time to avoid unwanted changes caused by text inputs. Our quantitative and qualitative evaluations reveal that our method can control generated images more faithfully to given texts than existing methods.

1 INTRODUCTION

Full-body human image synthesis holds great potential for content production and has been extensively studied in the fields of computer graphics and computer vision. In particular, recent advances in deep generative models have enabled us to create high-quality full-body human images. StyleGAN-Human (Fu et al., 2022) is a StyleGAN model (Karras et al., 2019; Karras et al., 2020) unsupervisedly trained using a large number of full-body human images. The users can instantly obtain realistic and diverse results from random latent codes, yet without intuitive control.

Text-based intuitive control of image synthesis has been an active research topic (Patashnik et al., 2021; Xia et al., 2021; Abdal et al., 2022; Wei et al., 2022; Kim et al., 2022; Gal et al., 2022; Wang et al., 2022; Ramesh et al., 2022) since the advent of CLIP (Radford et al., 2021), which learns cross-modal representations between images and texts. StyleCLIP (Patashnik et al., 2021) and HairCLIP (Wei et al., 2022) can control StyleGAN images by manipulating latent codes in accordance with given texts. These methods succeed in editing human and animal faces but struggle to handle full-body humans due to the much richer variations in garments and body shapes and poses. Specifically, these methods often neglect textual information on garments or deteriorate a person's identity (see Fig. 1).

In this paper, we propose a StyleGAN-based framework for text-based editing of garments in full-body human images, without sacrificing the person's identity. Our key insight is that the existing techniques of textual StyleGAN control have a problem with the latent code mapper, which manipulates StyleGAN latent codes according to input texts. Specifically, the modulation modules used in, e.g., HairCLIP's mapper equivalently modulate latent codes for StyleGAN layers and thus cannot identify and manipulate the text-specified latent codes. To address this issue, we present a latent code mapper architecture based on an attention mechanism, which can capture the correspondence between a given text and each latent code more accurately. In addition, we introduce feature-space masking at inference time to avoid unwanted changes in areas unrelated to input texts due to the latent code manipulation. This approach allows editing garments while preserving the person's identity. We demonstrate the effectiveness of our method through qualitative and quantitative comparisons with existing methods, including not only StyleGAN-based methods but also recent diffusion model-based methods.

[a] https://orcid.org/0000-0001-5043-8367
[b] https://orcid.org/0000-0001-5132-3350
[c] https://orcid.org/0000-0003-2843-1729

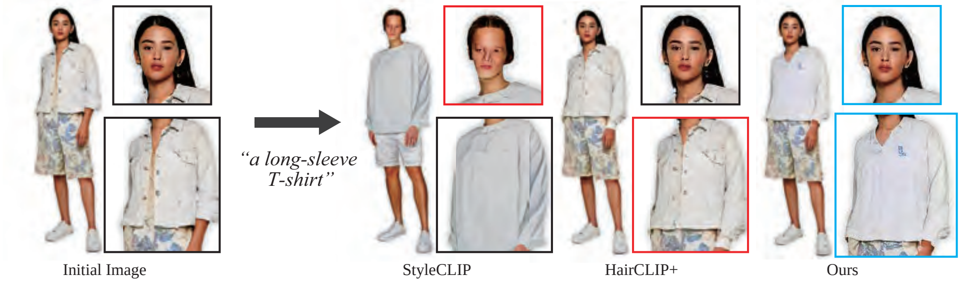

Figure 1: Garment editing comparison of existing methods and ours. StyleCLIP (Patashnik et al., 2021) erroneously changes the facial identity and pants. HairCLIP+ (a HairCLIP (Wei et al., 2022) variant trained with the same loss functions as ours) neglects the textual input due to its poor editing capability. Contrarily, our method successfully achieves virtual try-on of *"a long-sleeve T-shirt"* while preserving the facial identity and pants.

2 RELATED WORK

Generative adversarial networks. From the advent of generative adversarial networks (GANs) (Goodfellow et al., 2014), various studies have explored high-quality image synthesis by improving loss functions, learning algorithms, and network architectures (Arjovsky et al., 2017; Karras et al., 2018; Zhang et al., 2019; Brock et al., 2019). StyleGAN (Karras et al., 2019; Karras et al., 2020) is a milestone toward high-quality and high-resolution image synthesis. StyleGAN-Human (Fu et al., 2022) is a StyleGAN variant trained with an annotated full-body human image dataset. However, these unconditional models lack user controllability to generate images.

User-controllable image synthesis can be achieved via manipulation of latent codes in GANs. For example, unsupervised approaches (Chen et al., 2016; Voynov and Babenko, 2020; Härkönen et al., 2020; Shen and Zhou, 2021; He et al., 2021; Yüksel et al., 2021; Zhu et al., 2021; Oldfield et al., 2023) attempt to find interpretable directions in a latent space using, e.g., PCA and eigenvalue decomposition. However, finding desirable manipulation directions is not always possible. On the other hand, supervised approaches (Shen et al., 2020; Abdal et al., 2021; Yang et al., 2021; Jahanian et al., 2020; Spingarn et al., 2021) can manipulate latent codes to edit attributes corresponding to given annotations, such as gender and age. However, the manipulation is limited to specific attributes, and the annotation is costly. We thus leverage CLIP for text-based image manipulation without additional annotations.

Virtual try-on. Recently, 2D-based virtual try-on methods (Han et al., 2018; Wang et al., 2018; Yu et al., 2019; Song et al., 2020; Yang et al., 2020; Choi et al., 2021; Lee et al., 2022; Fele et al., 2022) have been actively studied. VTON (Han et al., 2018) and CP-VTON (Wang et al., 2018) are virtual try-on methods that learn the deformation and synthesis of garment images to fit target subjects. VTNFP (Yu et al., 2019) and ACGPN (Yang et al., 2020) synthesize images better preserving body and garment features by introducing a module that extracts segmentation maps. VITON-HD (Choi et al., 2021) and HR-VITON (Lee et al., 2022) allow virtual try-on for higher-resolution images. Although these methods require reference images of garment photographs, our method does not require reference images but uses texts as input guidance.

Text-guided image manipulation. There have been many studies on text-guided image manipulation (Patashnik et al., 2021; Xia et al., 2021; Abdal et al., 2022; Wei et al., 2022; Kim et al., 2022; Gal et al., 2022; Wang et al., 2022; Ramesh et al., 2022) by utilizing CLIP (Radford et al., 2021). StyleCLIP (Patashnik et al., 2021) proposes three methods (i.e., latent optimization, latent mapper, and global directions) to edit StyleGAN images using texts. In particular, the global direction method in $\mathcal{S}$ space (Wu et al., 2021) achieves fast inference while supporting arbitrary text input. HairCLIP (Wei et al., 2022) improved the StyleCLIP latent mapper to specialize in editing hairstyles using arbitrary text input. However, these methods focus on editing human and animal faces and are not suitable for full-body human images due to the much richer diversity in garments and body shapes and poses. These methods cannot appropriately reflect input texts to full-body human images and preserve the identity of face and body features.

Diffusion models for image generation and editing (Rombach et al., 2022; Kim et al., 2022; Couairon et al., 2022) have also attracted great attention. Recently, the diffusion model-based method specialized for fashion image editing (Baldrati et al., 2023) was proposed. These approaches provide high-quality editing but take several tens of times longer for infer-

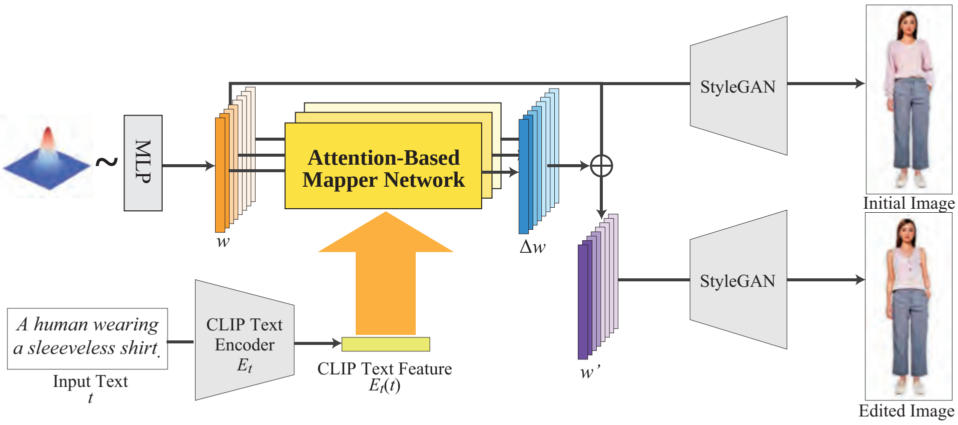

Figure 2: Overview of the proposed framework. The mapper network translates the latent codes w to the latent codes w' reflecting the text input. In the training time, only the mapper network is trained, and the other networks are freezed.

ence than StyleGAN-based methods. We also demonstrate that our method achieves higher-quality editing for full-body human images through comparisons with diffusion model-based methods in Section 4.

Very recently, FashionTex (Lin et al., 2023) was proposed to edit human images using texts and texture patches as input. Similar to our method, FashionTex also adopts latent code mappers for StyleGAN image manipulation, but our method differs from it in the following aspects. First, while FashionTex mainly aims to improve loss functions for existing latent code mappers, our focus is on extending the mapper architecture itself. Second, FashionTex needs reference texture patches to edit clothing textures, but our method uses only texts as input. Unfortunately, we cannot evaluate FashionTex because the complete source codes are not officially available yet. In the future, we would like to explore the potential of combining our method with FashionTex to leverage the advantages of each method.

3 PROPOSED METHOD

Fig. 2 illustrates an overview of the proposed framework. Inspired by HairCLIP (Wei et al., 2022), we adopt a latent code mapper trained to manipulate latent codes in the $\mathcal{W}+$ space of StyleGAN. The mapper network takes latent codes w and a text t as input and outputs the residual Δw between the input and edited latent codes. The input w is randomly sampled from Gaussian noise via the StyleGAN mapping network, and t is converted to a text feature $E_t(t)$ using the CLIP text encoder (Radford et al., 2021). Finally, we add Δw to w to create the edited latent code w', which is fed to the pre-trained StyleGAN to obtain an edited image. In the following sections, we describe the architecture of our latent code mapper (Section 3.1), training loss functions (Section 3.2), and feature-space masking in the StyleGAN generator (Section 3.3).

3.1 Mapper Network Architecture

The mapper network used in HairCLIP (Wei et al., 2022) has several blocks consisting of a fully connected layer, modulation module, and activation function. The modulation module modulates latent code features normalized through a LayerNorm layer using the scaling and shifting parameters f_γ and f_β computed from CLIP text features (see the bottom left diagram in Fig. 3). HairCLIP uses three mappers (coarse, medium, and fine) to handle different semantic levels of a latent code fed to each StyleGAN layer. However, the modulation modules in each mapper equivalently modulate given latent codes. Therefore, each mapper cannot identify and manipulate only latent codes related to input texts. As a result, the HairCLIP mapper cannot reflect input texts well for full-body human images.

To manipulate appropriate latent codes according to text input, we introduce a cross-attention mechanism into our latent code mapper. Fig. 3 shows our network architecture. Our network first applies positional encoding to distinguish between latent codes fed to different StyleGAN layers. Then, we apply the modulation module used in HairCLIP which uses the CLIP text features $E_t(t)$ to modulate the intermediate output. In addition, following the Transformer architecture (Vaswani et al., 2017), we adopt the multi-head cross-attention mechanism, which can capture multiple relationships between input features. To compute the multi-head cross attention, we define

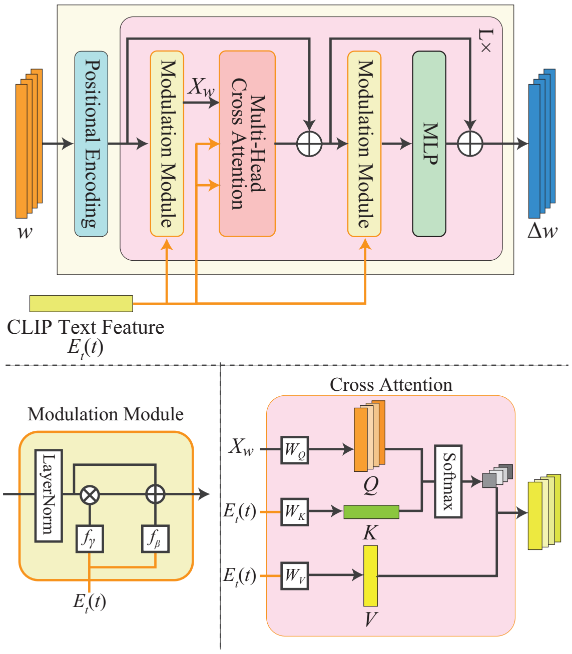

Figure 3: Architecture of our latent code mapper (top). Given latent codes w and a CLIP text feature $E_t(t)$, it estimates the residual Δw between input and edited latent codes. The latent codes are manipulated according to an input text via the cross-attention mechanism (bottom right) besides the HairCLIP (Wei et al., 2022) modulation module (bottom left).

the query Q, key K, and value V as follows:

$$Q = X_w W_Q, \quad K = E_t(t) W_K, \quad V = E_t(t) W_V, \quad (1)$$

where the query Q is computed from the latent code feature $X_w \in \mathbb{R}^{N \times 512}$ (N is the number of StyleGAN layers taking latent codes), and the key K and value V are computed from the CLIP feature $E_t(t) \in \mathbb{R}^{1 \times 512}$ of the input text t. The tensors $W_Q, W_K, W_V \in \mathbb{R}^{512 \times 512}$ are the weights to be multiplied with each input. Using the query Q_i, key K_i, and value V_i for a head i, the multi-head cross attention is defined as:

$$MultiHead(Q, K, V) = [Softmax(\frac{Q_i K_i^T}{\sqrt{d}}) V_i]_{i=1:h} W_o, \quad (2)$$

where $d = 512/h$ (h is the number of heads), and $W_o \in \mathbb{R}^{512 \times 512}$ is the weight to be multiplied with the concatenated attentions of the multiple heads. Note that, unlike the typical multi-head cross attention, our method applies the softmax function along the column direction to ensure that the weights for all latent code features sum to 1. We repeat the block consisting of the modulation modules, multi-head cross attention, and multilayer perceptron (MLP) L times, as illustrated in Fig. 3.

3.2 Loss Functions

In the mapper network, we aim to acquire latent codes capable of generating images reflecting the input text while preserving unrelated areas. We first adopt the CLIP loss following the approach of StyleCLIP (Patashnik et al., 2021).

$$\mathcal{L}_{clip} = 1 - \cos(E_i(G(w')), E_t(t)), \quad (3)$$

where $\cos(\cdot,\cdot)$ denotes the cosine similarity, E_i and E_t are the image and text encoders of CLIP, respectively, and $G(w')$ is the image generated from the edited latent code w'. In addition, we introduce the directional CLIP loss presented in StyleGAN-NADA (Gal et al., 2022).

$$\mathcal{L}_{direct} = 1 - \frac{\Delta T \cdot \Delta I}{\|\Delta T\| \, \|\Delta I\|}, \quad (4)$$

where $\Delta T = E_t(t) - E_t(t_{source})$ and $\Delta I = E_i(G(w')) - E_i(G(w))$. One of the purposes of the directional CLIP loss in StyleGAN-NADA is to finetune the StyleGAN to avoid mode collapse caused by the CLIP loss. Meanwhile, our method does not finetune StyleGAN, but the directional CLIP loss encourages the mapper not to train many-to-one mapping between latent codes and has an important role in generating diverse results. Besides, we define the background loss so that areas unrelated to texts do not change:

$$\mathcal{L}_{bg} = \left\| (\bar{P}_t(G(w)) \cap \bar{P}_t(G(w'))) * (G(w) - G(w')) \right\|_2, \quad (5)$$

where $\bar{P}_t(G(w))$ is the binary mask representing the outside of target garment areas extracted using the off-the-shelf human parsing model (Li et al., 2020), and $*$ denotes element-wise multiplication. Finally, to maintain the quality of the generated image, we introduce the L2 regularization for the residual of latent codes Δw.

$$\mathcal{L}_{norm} = \|\Delta w\|_2 . \quad (6)$$

The final loss $\mathcal{L}_{final}$ is defined as:

$$\mathcal{L}_{final} = \lambda_c \, \mathcal{L}_{clip} + \lambda_d \, \mathcal{L}_{direct} + \lambda_b \, \mathcal{L}_{bg} + \lambda_n \, \mathcal{L}_{norm}, \quad (7)$$

where $\lambda_c, \lambda_d, \lambda_b$, and λ_n are the weights for corresponding loss functions.

3.3 Feature-space Masking

Although the background loss (Eq. (5)) restricts changes in unrelated areas to some extent, it is insufficient due to the limited controllability in the low-dimensional latent space. Therefore, we further restrict editable areas using feature-space masking, inspired by the approach by Jakoel et al. (Jakoel et al., 2022). However, unlike their user-specified static

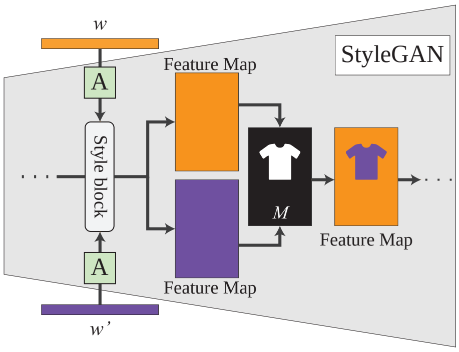

Figure 4: Overview of feature-space masking. Given a mask M, we merge two feature maps computed using latent codes w and w' in each style block (Karras et al., 2020).

masking, we have to handle masks whose shapes change dynamically according to input texts. Furthermore, there is a chicken-and-egg problem; we require a mask to generate an output image, whereas we require the output image to determine the mask shape. We solve this problem as follows. First, we generate images $G(w)$ and $G(w')$ without masking using the input latent code w and the edited latent code w'. Second, we apply the human parsing network (Li et al., 2020) to obtain binary masks $P_t(G(w))$ and $P_t(G(w'))$ of the target garment. Finally, we merge both masks because, in case that the edited garment is smaller than the original, the original garment appears in the final image:

$$M = P_t(G(w)) \cup P_t(G(w')). \tag{8}$$

Using this mask M, we modify a part of the StyleGAN's convolution layers and combine two feature maps created from latent codes w and w' during inference, as shown in Fig. 4. By merging an input image and an edited result in the feature space, we can obtain more natural results than pixel-space masking, as discussed in Section 4.2.

4 EXPERIMENTS

Implementation details. We implemented our method using Python and PyTorch, and ran our program on NVIDIA Quadro RTX 6000. It took about 0.3 seconds to obtain an edited image. The dataset contains 30,000 images synthesized with StyleGAN-Human (Fu et al., 2022) from random latent codes. We used 28,000 sets for training and 2,000 for testing, in which each set contains an image and the corresponding latent code for each layer. For the text input, we prepared 10 text descriptions of upper-body garment shapes, 16 text descriptions of lower-body garment shapes, and 15 text descriptions of garment textures. To help our latent code mappers learn disentangled garment editing, we trained the mapper networks separately for the upper and lower bodies. The mappers were trained using the pairs of training latent codes and a random text description corresponding to each body part. Following HairCLIP (Wei et al., 2022), we divided the latent codes into three groups (coarse, medium, and fine) and prepared a mapper network for each group. We created separate mapper networks for the upper and lower body to facilitate effective training. Appendix provides more details about the training configurations.

Compared methods. We compared our method with existing StyleGAN-based methods and diffusion model-based methods. For the StyleGAN-based methods, we used StyleCLIP (Patashnik et al., 2021) and HairCLIP (Wei et al., 2022) combined with StyleGAN-Human (Fu et al., 2022). For StyleCLIP, we used the global direction method in $\mathcal{S}$ space (Wu et al., 2021) among the three proposed methods because it is fast and can handle arbitrary texts. To adapt HairCLIP to full-body human images, we changed the original loss functions designed for editing hairstyles to the same loss functions as our method. We denote this modified method as HairCLIP+. For diffusion model-based methods, we used Stable Diffusion-based inpainting (SD inpainting) (Rombach et al., 2022) and DiffEdit (Couairon et al., 2022). Because SD inpainting requires masks of inpainted regions, we created them using the off-the-shelf human parsing model (Li et al., 2020). Meanwhile, DiffEdit can automatically estimate mask regions related to text inputs and edit those regions. Details on the implementation of each method are provided in Appendix.

Evaluation metrics. As the objective evaluation metrics for quantitative comparison, we used CLIP Acc and BG LPIPS. CLIP Acc evaluates whether edited images reflect the semantics of input texts. Inspired by the work by Parmar et al. (Parmar et al., 2023), we define CLIP Acc as the percentage of instances (i.e., test images) where the target text has a higher CLIP similarity (Radford et al., 2021) to the edited image than the input image. BG LPIPS evaluates the preservation degree of background regions outside target garment areas. We calculated LPIPS (Zhang et al., 2018) between masked areas of the input and edited images. The masks are extracted using the off-the-shelf human parsing model (Li et al., 2020). We computed CLIP Acc and BG LPIPS for 2,000 test images, which were edited using text inputs randomly selected from the prepared text descriptions.

Table 1: Quantitative comparison with the existing methods, StyleCLIP (Patashnik et al., 2021), and HairCLIP+ (Wei et al., 2022). The bold and underlined values show the best and second best scores.

Method	CLIP Acc ↑	BG LPIPS ↓
StyleCLIP	**98.0%**	0.204
HairCLIP+	80.5%	**0.028**
Ours w/o masking	97.9%	0.075

Table 2: Quantitative comparison with the existing methods, StyleCLIP (Patashnik et al., 2021), and HairCLIP+ (Wei et al., 2022), with our feature-space masking.

Method	CLIP Acc ↑	BG LPIPS ↓
StyleCLIP w/ masking	77.6%	0.027
HairCLIP+ w/ masking	61.1%	**0.004**
Ours w/ masking	**82.2%**	0.016

4.1 Evaluating Latent Code Mapper

We first evaluate the effectiveness of our latent code mapper without our feature-space masking. As shown in Table 1, StyleCLIP has the best score in CLIP Acc but the significantly worst score in BG LPIPS. The qualitative results in Fig. 7 also show that StyleCLIP changed the facial identity and garments unrelated to the text input. In contrast, our method has overall good scores in both metrics, which means that the edited results faithfully follow the text input while preserving unrelated areas. Finally, HairCLIP+ has the worst score in CLIP Acc, although it used the same loss functions as ours. In other words, our mapper more effectively learned text-based latent code transformation than the HairCLIP mapper in the domain of full-body human images.

4.2 Evaluating Feature-space Masking

We evaluated the effectiveness of our feature-space masking. First, we compared our feature-space masking with pixel-space masking, which merges target areas of edited images and the other regions of the input images in the pixel space. As shown in Fig. 5, pixel-space masking yields unnatural results containing artifacts around the boundaries of garments. In contrast, feature-space masking obtains plausible results without such artifacts.

Next, we applied feature-space masking to StyleCLIP, HairCLIP+, and our method. As can be seen in Fig. 6, feature-space masking enables the existing methods to preserve areas unrelated to the specified text description, but the text input is not reflected in the outputs appropriately. In addition, the quantitative comparisons in Tables 1 and 2 show that feature-space masking significantly drops CLIP Acc for StyleCLIP and HairCLIP+. These performance drops come from the fact that the existing methods improve CLIP Acc by manipulating background regions rather than target garment regions. In contrast, thanks to our latent code mapper, which can reflect textual information to appropriate latent codes for editing target regions, our method with feature-space masking shows the best CLIP Acc while improving BG LPIPS.

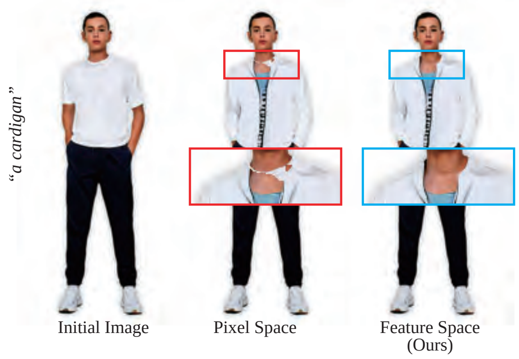

Figure 5: Qualitative comparison of pixel-space masking and feature-space masking.

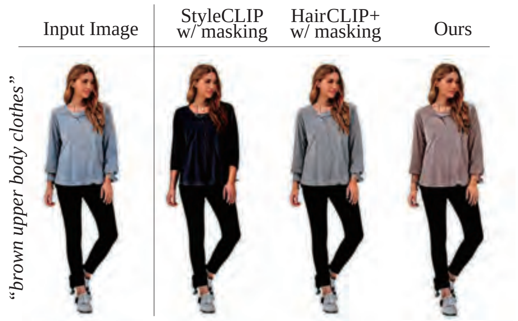

Figure 6: Qualitative comparison with the existing methods, StyleCLIP (Patashnik et al., 2021), and HairCLIP+ (Wei et al., 2022), with feature-space masking.

4.3 Comparison with Existing Methods

Fig. 7 shows the qualitative comparison between our method with feature-space masking and the existing methods. Some results of SD Inpainting and DiffEdit effectively reflect the input text information but contain artifacts and lose fine details of faces and hands. The results of StyleCLIP in the first row show that the garment textures change together with the garment shape, even though the input text is specified to edit the shape only. In addition, the results from the second row show that StyleCLIP struggles to edit the garment textures according to the input texts. HairCLIP+ often outputs results that hardly follow the input texts. In this case, the latent code mapper of HairCLIP for face images cannot be adapted to full-body human images well. In contrast, our method correctly reflects the text semantics in the output images while preserv-

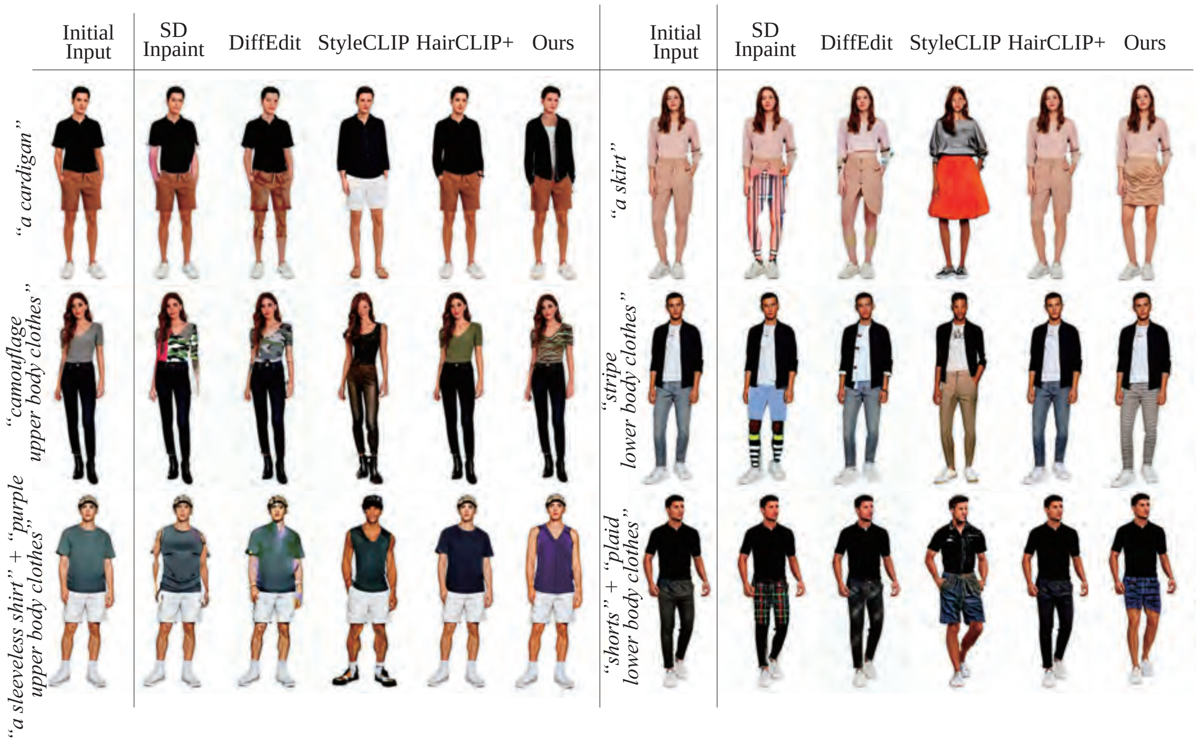
Initial Input
SD Inpaint
DiffEdit
StyleCLIP
HairCLIP+
Ours
Initial Input
SD Inpaint
DiffEdit
StyleCLIP
HairCLIP+
Ours
"a cardigan"
"camouflage upper body clothes"
"a sleeveless shirt" + "purple upper body clothes"
"a skirt"
"stripe lower body clothes"
"shorts" + "plaid lower body clothes"

Figure 7: Qualitative comparison with the existing methods (Rombach et al., 2022; Couairon et al., 2022; Patashnik et al., 2021; Wei et al., 2022).

Table 3: User study results. Users were asked to rate alignment to text and realism of images generated by each method.

Method	Text alignment ↑	Realism ↑
SD Inpainting	2.42	2.24
DiffEdit	2.10	2.42
StyleCLIP	2.75	2.84
HairCLIP+	2.50	**4.29**
Ours	**3.50**	4.06

ing the unrelated areas. Regarding the computational time for generating a single image, the StyleGAN-based methods (i.e., StyleCLIP and HairCLIP) took approximately 0.1 to 0.5 seconds, while SD Inpaint and DiffEdit took roughly 2 and 10 seconds, respectively. Please refer to Appendix for more results.

User study. We conducted a subjective user study to validate the effectiveness of our method. We asked 13 participants to evaluate 20 random sets of images edited using our method and the compared methods. The participants scored the edited images on a 5-point scale in terms of text alignment and realism. Table 3 shows the average scores for each method. Our method obtains the best score for text alignment and is on par with HairCLIP+ for realism.

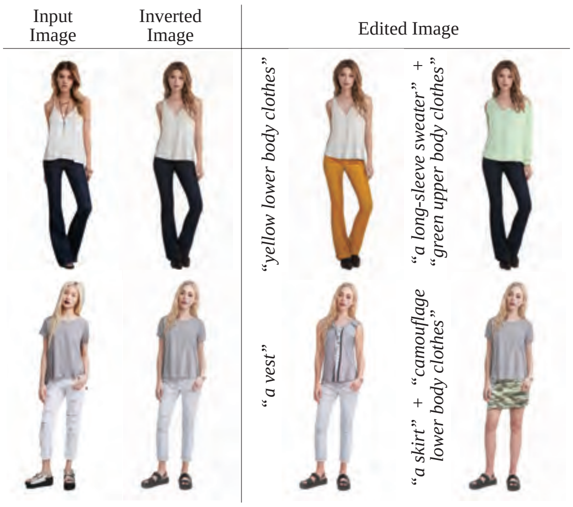
Input Image
Inverted Image
Edited Image
"yellow lower body clothes"
"a long-sleeve sweater" + "green upper body clothes"
"a vest"
"a skirt" + "camouflage lower body clothes"

Figure 8: Application to real images.

4.4 Application

We also validate the effectiveness of our method for real images. We used e4e (Tov et al., 2021) to invert real images to latent codes and fed them to our mapper network. We trained the e4e encoder on the SHHQ dataset containing 256×512 images collected for StyleGAN-Human (Fu et al., 2022). For training

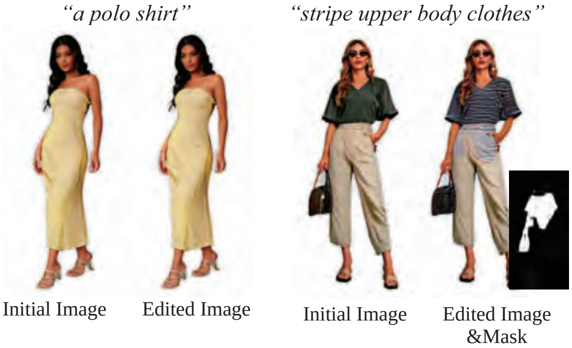

Figure 9: Failure cases. Our method cannot handle full-body garments like a dress (left). In addition, inaccurate masks estimated by the human parsing model change unintended areas (right).

the e4e encoder, we used the official default parameters, with an only modification to set the ID loss weight to zero because the ID loss is defined only for faces. As shown in Fig. 8, our method can edit real images accroding to given texts. Although the inverted images lose the details of the faces and shoes, this problem arises from GAN inversion and can be alleviated by improving the inversion method in the future.

5 CONCLUSIONS

In this paper, we tackled a problem of controlling StyleGAN-Human using text input. To this end, we proposed a mapper network based on an attention mechanism that can manipulate appropriate latent codes according to text input. In addition, we introduced feature-space masking at inference time to improve the performance of identity preservation outside target editing areas. Qualitative and quantitative evaluations demonstrate that our method outperforms existing methods in terms of text alignment, realism, and identity preservation.

Limitations and future work. Currently, our mapper networks are trained separately for the upper and lower bodies. The user needs to select the mapper networks depending on the target texts. In addition, we cannot handle full-body garments like a dress (see the left side of Fig. 9). In the future, we want to develop a method to automatically determine which body parts should be edited according to text inputs. In addition, as shown in the right side of Fig. 9, our method sometimes changes unintended areas depending on the mask M's accuracy. This problem could be improved using more accurate human parsing models.

REFERENCES

Abdal, R., Zhu, P., Femiani, J., Mitra, N., and Wonka, P. (2022). Clip2StyleGAN: Unsupervised extraction of StyleGAN edit directions. In *ACM SIGGRAPH conference proceedings*, pages 1–9.

Abdal, R., Zhu, P., Mitra, N. J., and Wonka, P. (2021). StyleFlow: Attribute-conditioned exploration of StyleGAN-generated images using conditional continuous normalizing flows. *ACM Trans. Graph.*, 40(3):21:1–21:21.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *ICML*, pages 214–223.

Baldrati, A., Morelli, D., Cartella, G., Cornia, M., Bertini, M., and Cucchiara, R. (2023). Multimodal garment designer: Human-centric latent diffusion models for fashion image editing. In *ICCV*.

Brock, A., Donahue, J., and Simonyan, K. (2019). Large scale GAN training for high fidelity natural image synthesis. In *ICLR*.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. (2016). InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *NeurIPS*, pages 2172–2180.

Choi, S., Park, S., Lee, M., and Choo, J. (2021). VITON-HD: High-resolution virtual try-on via misalignment-aware normalization. In *CVPR*, pages 14131–14140.

Couairon, G., Verbeek, J., Schwenk, H., and Cord, M. (2022). Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*.

Fele, B., Lampe, A., Peer, P., and Struc, V. (2022). C-VTON: Context-driven image-based virtual try-on network. In *WACV*, pages 3144–3153.

Fu, J., Li, S., Jiang, Y., Lin, K., Qian, C., Loy, C. C., Wu, W., and Liu, Z. (2022). StyleGAN-Human: A data-centric odyssey of human generation. *CoRR*, abs/2204.11823.

Gal, R., Patashnik, O., Maron, H., Bermano, A. H., Chechik, G., and Cohen-Or, D. (2022). StyleGAN-NADA: CLIP-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *NeurIPS*.

Han, X., Wu, Z., Wu, Z., Yu, R., and Davis, L. S. (2018). Viton: An image-based virtual try-on network. In *CVPR*.

Härkönen, E. et al. (2020). GANSpace: Discovering interpretable GAN controls. In *NeurIPS*.

He, Z., Kan, M., and Shan, S. (2021). EigenGAN: Layer-wise eigen-learning for GANs. In *ICCV*.

Jahanian, A., Chai, L., and Isola, P. (2020). On the "steerability" of generative adversarial networks. In *ICLR*.

Jakoel, K., Efraim, L., and Shaham, T. R. (2022). GANs spatial control via inference-time adaptive normalization. In *WACV*, pages 2160–2169.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2018). Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*.

Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and improving the image quality of StyleGAN. In *CVPR*, pages 8110–8119.

Kim, G., Kwon, T., and Ye, J. C. (2022). DiffusionCLIP: Text-guided diffusion models for robust image manipulation. In *CVPR*, pages 2426–2435.

Lee, S., Gu, G., Park, S., Choi, S., and Choo, J. (2022). High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *ECCV*, pages 204–219.

Li, P., Xu, Y., Wei, Y., and Yang, Y. (2020). Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3260–3271.

Lin, A., Zhao, N., Ning, S., Qiu, Y., Wang, B., and Han, X. (2023). Fashiontex: Controllable virtual try-on with text and texture. In *ACM SIGGRAPH Conference Proceedings*, pages 56:1–56:9. ACM.

Oldfield, J., Tzelepis, C., Panagakis, Y., Nicolaou, M. A., and Patras, I. (2023). PandA: Unsupervised learning of parts and appearances in the feature maps of GANs. In *ICLR*.

Parmar, G., Singh, K. K., Zhang, R., Li, Y., Lu, J., and Zhu, J. (2023). Zero-shot image-to-image translation. *CoRR*, abs/2302.03027.

Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., and Lischinski, D. (2021). StyleCLIP: Text-driven manipulation of stylegan imagery. In *ICCV*, pages 2085–2094.

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 *ICML*, pages 8748–8763.

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*.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695.

Shen, Y., Gu, J., Tang, X., and Zhou, B. (2020). Interpreting the latent space of GANs for semantic face editing. In *CVPR*, pages 9240–9249.

Shen, Y. and Zhou, B. (2021). Closed-form factorization of latent semantics in GANs. In *CVPR*, pages 1532–1540.

Song, D., Li, T., Mao, Z., and Liu, A.-A. (2020). SP-VITON: shape-preserving image-based virtual try-on network. *Multimedia Tools and Applications*, 79:33757–33769.

Spingarn, N., Banner, R., and Michaeli, T. (2021). GAN "steerability" without optimization. In *ICLR*.

Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., and Cohen-Or, D. (2021). Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *NeurIPS*.

Voynov, A. and Babenko, A. (2020). Unsupervised discovery of interpretable directions in the GAN latent space. In *ICML*, pages 9786–9796.

Wang, B., Zheng, H., Liang, X., Chen, Y., Lin, L., and Yang, M. (2018). Toward characteristic-preserving image-based virtual try-on network. In *ECCV*, pages 589–604.

Wang, H., Lin, G., del Molino, A. G., Wang, A., Yuan, Z., Miao, C., and Feng, J. (2022). ManiCLIP: Multi-attribute face manipulation from text. *arXiv preprint arXiv:2210.00445*.

Wei, T., Chen, D., Zhou, W., Liao, J., Tan, Z., Yuan, L., Zhang, W., and Yu, N. (2022). HairCLIP: Design your hair by text and reference image. In *CVPR*, pages 18072–18081.

Wright, L. (2019). Ranger - a synergistic optimizer. https://github.com/lessw2020/Ranger-Deep-Learning-Optimizer.

Wu, Z., Lischinski, D., and Shechtman, E. (2021). Stylespace analysis: Disentangled controls for StyleGAN image generation. In *CVPR*, pages 12863–12872.

Xia, W., Yang, Y., Xue, J.-H., and Wu, B. (2021). TediGAN: Text-guided diverse face image generation and manipulation. In *CVPR*, pages 2256–2265.

Yang, H., Chai, L., Wen, Q., Zhao, S., Sun, Z., and He, S. (2021). Discovering interpretable latent space directions of GANs beyond binary attributes. In *CVPR*, pages 12177–12185.

Yang, H., Zhang, R., Guo, X., Liu, W., Zuo, W., and Luo, P. (2020). Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *CVPR*, pages 7850–7859.

Yu, R., Wang, X., and Xie, X. (2019). VTNFP: An image-based virtual try-on network with body and clothing feature preservation. In *ICCV*, pages 10511–10520.

Yüksel, O. K., Simsar, E., Er, E. G., and Yanardag, P. (2021). LatentCLR: A contrastive learning approach for unsupervised discovery of interpretable directions. In *ICCV*, pages 14243–14252.

Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. (2019). Self-attention generative adversarial networks. In *ICML*, pages 7354–7363.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*.

Zhu, J., Feng, R., Shen, Y., Zhao, D., Zha, Z., Zhou, J., and Chen, Q. (2021). Low-rank subspaces in GANs. In *NeurIPS*.

APPENDIX

Hyperparameters. Our method used the pre-trained StyleGAN-Human (Fu et al., 2022) model, which has the structure of StyleGAN2 (Karras et al., 2020) with a modification to output 256×512 images. We used a truncation value of $\psi = 0.7$ to generate images for training and testing. The StyleGAN-Human model consists of a total of 16 layers, which are divided into three stages (i.e., course, middle, fine) with 4, 4, and 8 layers, respectively. For our mapper network (see Section 3.1), we set the internal block repetition count L (see also Figure 3) to 6 and the number of heads h of the multi-head cross attention (Vaswani et al., 2017) to 4. The loss weights $\lambda_c, \lambda_d, \lambda_b$, and λ_n were set to 1.0, 2.0, 5.0, and 1.0, respectively. We employed the Ranger (Wright, 2019) optimizer with a learning rate of 0.0005 and $(\beta_1, \beta_2) = (0.95, 0.9)$.

Implementation of existing methods. For StyleCLIP (Patashnik et al., 2021) and HairCLIP (Wei et al., 2022), we used the official implementations[1][2] with a modification to replace StyleGAN with StyleGAN-Human, and reran the preprocessing and training. For Stable Diffusion-based inpainting (SD inpainting) (Rombach et al., 2022) and DiffEdit (Couairon et al., 2022), we used the Stable Diffusion version 1.4. For SD inpainting, we used the image generation pipeline of the Diffusers library[3]. For DiffEdit, we used the unofficial implementation[4] because no official implementation has been released.

Input texts. We synthesized input texts for training by inserting labels into text templates. Table 4 shows the list of labels. For input text templates, we adopted "*a human wearing* {*shape label*} " for shape manipulation and "*a human wearing* {*texture label*} *upper body (lower body) clothes*" for texture manipulation. For texture manipulation, we randomly picked a label from the same texture label list for both upper and lower bodies. The input t_{source} of the directional CLIP loss is set to "*a human*".

Creating masks using a human parsing model. In our method, we use the off-the-shelf human parsing model (Li et al., 2020) to create masks for loss calculation during training and feature-space masking during inference. The human parsing model segments a full-body human image into 18 semantic regions. We create masks by selecting specific semantic regions, which differ depending on the editing areas (i.e., upper body or lower body) and the types of editing (i.e., shape or texture). Table 5 shows the selected semantic regions in each case.

Table 4: Label list for training.

Shape of upper body clothes	Shape of lower body clothes	Texture
a sleeveless shirt a long-sleeve sweater a long-sleeve T-shirt a hoodie a cardigan a dress shirt a polo shirt a denim shirt a jacket a vest	pants slacks dress pants jeans shorts cargo pants capri pants cropped pants chino pants leggings wide pants a jogger a skirt a miniskirt a long skirt a tight skirt	purple red orange yellow green blue gray brown black white pink stripes dots plaid camouflage

Table 5: Selected semantic regions for mask creation.

	Upper body	Lower body
Shape	Upper-clothes Left-arm Right-arm	Skirt Pants Left-leg Right-leg
Texture	Upper-clothes	Skirt Pants

Additional qualitative comparison. Figures 10 and 11 show the additional qualitative comparisons. Some results of SD Inpainting and DiffEdit effectively reflect the input text information but contain artifacts and lose fine details of faces and hands. The results of StyleCLIP in the first row in Fig. 10 show that the garment textures change together with the garment shape, even though the input text is specified to edit the shape only. In addition, the results from the third and fourth rows in Figures 10 and 11 show that StyleCLIP struggles to edit the garment textures according to the input texts. HairCLIP+ often outputs results that hardly follow the input texts. In this case, the latent code mapper of HairCLIP for face images cannot be adapted to full-body human images well. In contrast, our method correctly reflects the text semantics in the output images while preserving the unrelated areas.

[1] https://github.com/orpatashnik/StyleCLIP

[2] https://github.com/wty-ustc/HairCLIP

[3] https://github.com/huggingface/diffusers

[4] https://github.com/Xiang-cd/DiffEdit-stable-diffusion/

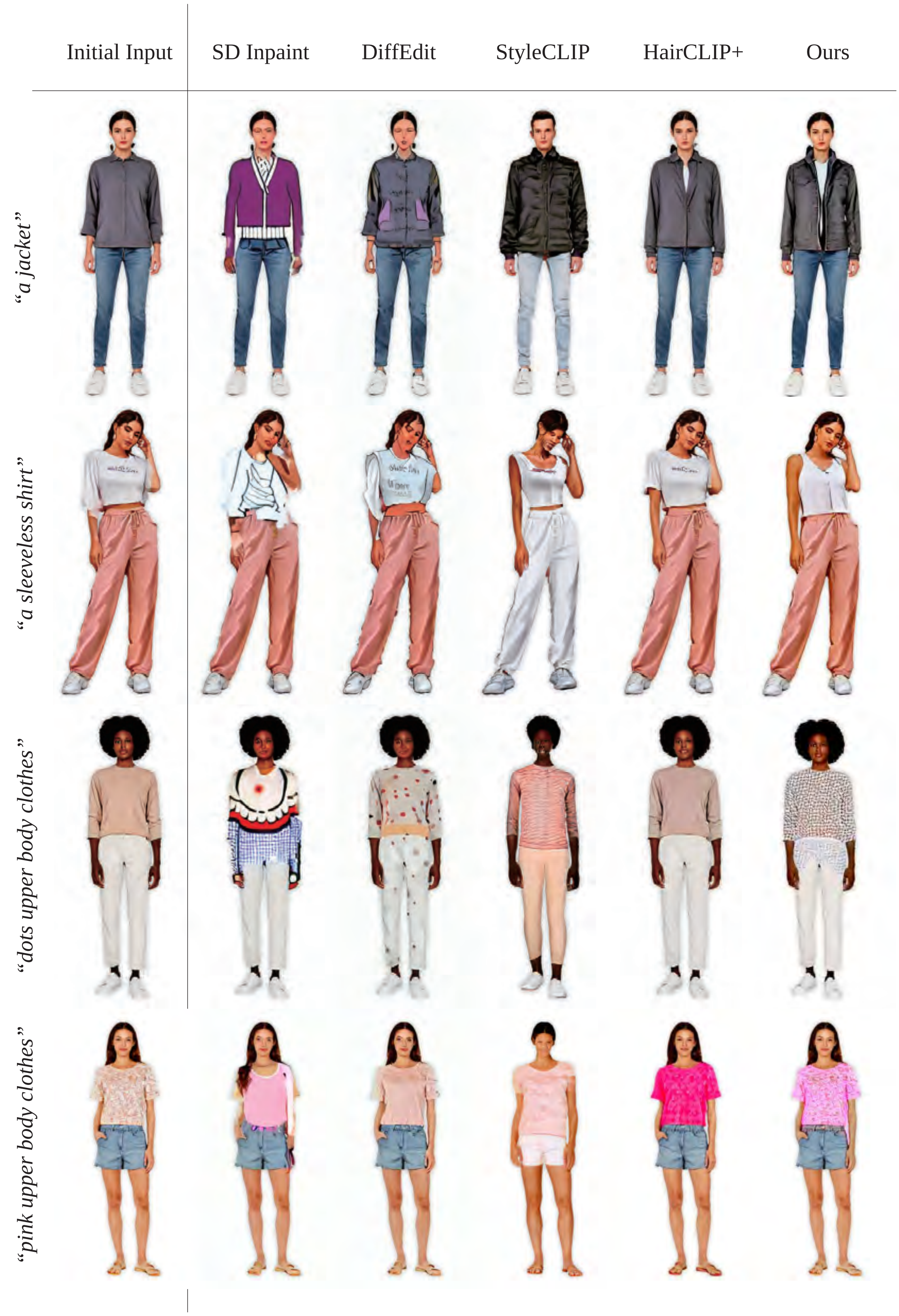

Figure 10: Additional qualitative comparison for upper body clothes manipulation.

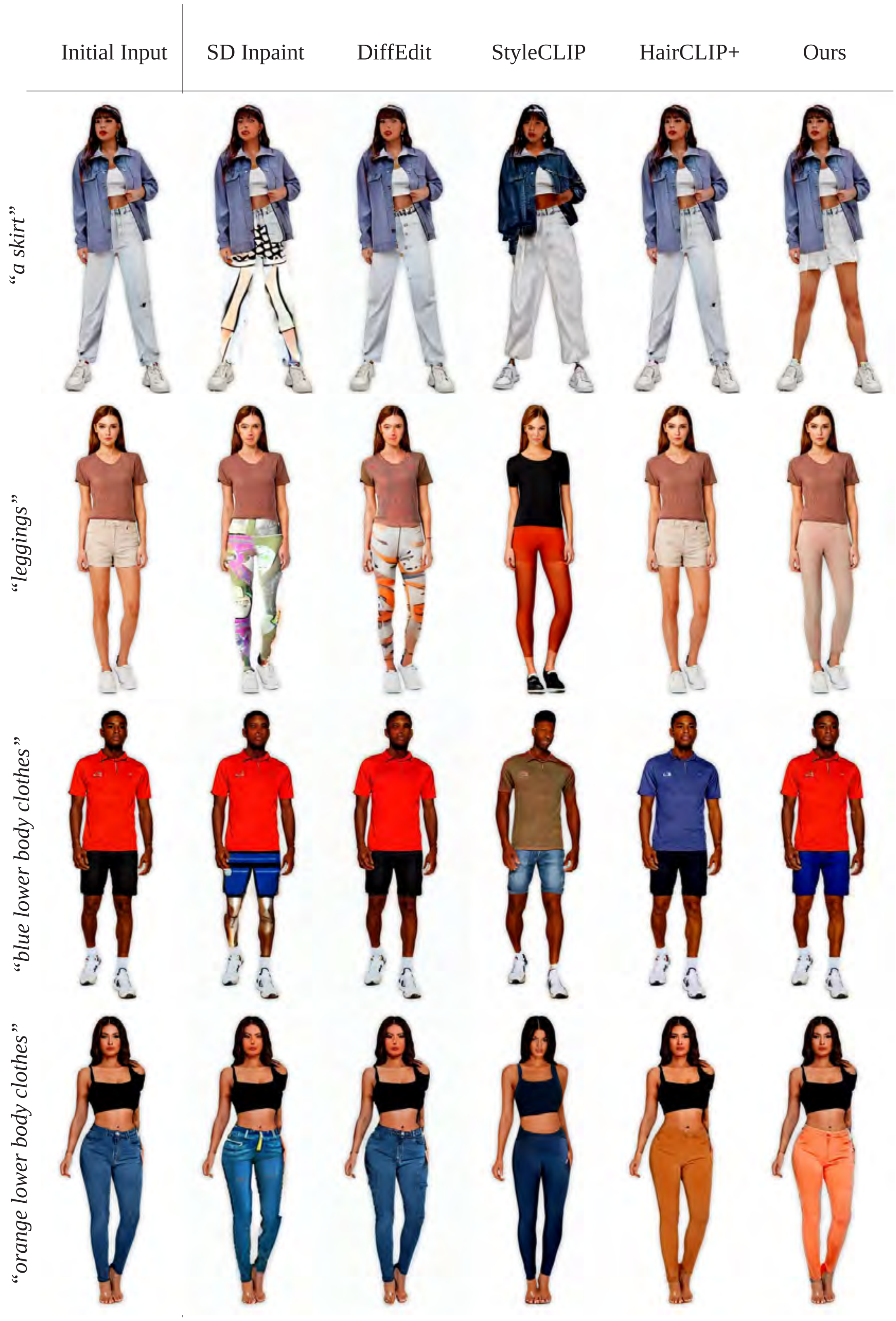

Figure 11: Additional qualitative comparison for lower body clothes manipulation.