FashionTailor: Controllable Clothing Editing for Human Images with Appearance Preserving

Jie Hou[1], Jianghong Ma[1], Xiangyu Mu[1], Haijun Zhang[1]*, Zhao Zhang[2]

[1]Department of Computer Science, Harbin Institute of Technology, Shenzhen 518055, China
[2]Department of Computer Science, Hefei University of Technology, Hefei 230009, China
arlo@stu.hit.edu.cn, majianghong@hit.edu.cn, 21B951013@stu.hit.edu.cn, hjzhang@hit.edu.cn, cszzhang@gmail.com

Abstract

The garment structure serves as a crucial medium for expressing the designer's creative vision and showcasing the distinctive character of clothing items. Effective editing of garment structure in fashion images allows for an advanced preview of the design, accelerating the process of garment customization to meet individualized requirements. Although large-scale diffusion models have demonstrated impressive image generation and editing capabilities, no efforts have been made to exploit their potential in part-level editing of images. Unlike previous research, we define a clothing structure editing (CSE) task aimed at accurately editing the local structure of human-centered clothing images through simple instruction-based prompts while maintaining the consistency of clothing appearance. Specifically, this paper develops a new controllable triple-flow framework for structure editing named Fashion-Tailor. An additional network called ClothingNet is proposed to extract the clothing details to address the rigid constraints of the original garment structure. Then, we propose a semantic-refined module to extract the semantic understanding of the source image and adaptively focus on the part to be edited. We also design a cross-blend attention mechanism to integrate fine-grained clothing features to guarantee precise alignment between appearance and target structure features. In addition, a garment structure dataset called StructureFashion has been collated, wherein each item of clothing is represented by multiple photos with diverse structure characteristics, containing over six million pairs. Finally, our method supports editing the structure of multiple parts on a garment simultaneously. Extensive experiments validate the effectiveness of our meth-od for editing part-level human images in StructureFashion dataset and real-scenarios.

Figure 1: **Structure editing results** generated by FashionTailor, which enables the editing of multiple parts of arbitrary garments simultaneously. Notably, the source images for rows 2-4 are from the **wild dataset**.

Introduction

The rise of e-commerce and artificial intelligence is transforming how consumers purchase clothing, making bespoke clothing production increasingly popular (Nobile and Cantoni 2023). However, the bespoke clothing process is usually complex and challenging for both consumers and merchants. Consumers often need design alterations, reworks, and even chargebacks due to the difficulty in visualizing how a garment looks when worn. For merchants, hiring models to photograph garments with the same style but different shapes is a time-consuming and labor-intensive process. Thus, obtaining a preview of garment alterations dressed on models becomes essential. This necessitates a tool capable of directly modifying the clothing structure within human images. Clothing structure (Chen 2020) encompasses the design elements of clothing products, including the combina-

*Corresponding author.

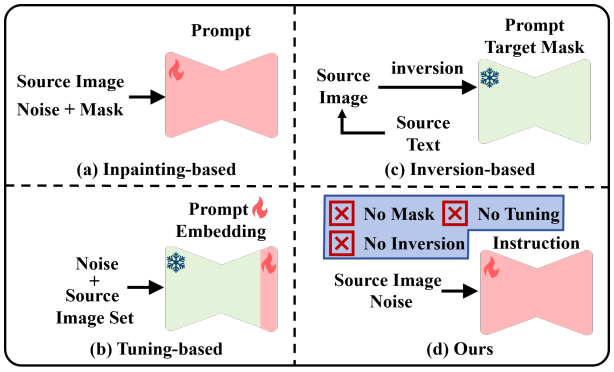

Figure 2: Comparisons between different image editing pipelines. Our pipeline follows the idea of IP2P and requires only a simple instruction for part-level editing.

tion of clothing parts such as neckline, length, and sleeve shape, which determine the overall shape, silhouette, and dressing effect (Hu 2023). Different combinations of parts create unique fashion styles and appropriate dressing scenarios. Consequently, we define a ***Clothing Structure Editing (CSE)*** task with broad applications in virtual reality, fashion e-commerce, and photo beautification. CSE technology focuses on accurately performing editing instructions while maintaining the clothing's texture characteristics, as shown in Figure 1. Unlike object-level local editing (Duan et al. 2024), the CSE task involves part-level local manipulation, which is more challenging due to the interdependence of local parts and the difficulty in isolating them from the whole.

Potential Works. Some early works based on generative adversarial network (GAN) (Goodfellow et al. 2020) can be extended to solve the CSE task (Zhang et al. 2021; Xie et al. 2022). These methods typically perform human image synthesis and editing by decoupling human attributes (e.g., pose) from clothing appearance attributes (e.g., texture and shape). Obviously, these works have two limitations. First, they require an exact semantic graph input or are limited to length editing. Second, GAN-based methods may suffer from unstable training (Gulrajani et al. 2017) and modal collapse (Miyato et al. 2018), leading to losing image details.

Our Work. In this paper, we introduce ***FashionTailor***, a new diffusion-based method designed to enable user-friendly, efficient, and precise control of CSE, much like the work of a skilled tailor. Unlike current local image editing paradigms, as shown in Figure 2, we develop a triple-flow framework consisting of appearance, structure, and denoising flow. **First**, the appearance flow processes clothing texture, skin features, and background information. In contrast, the structure flow processes the target structural information through human-written instructions, facilitating intuitive image editing without requiring explicit masks. Inspired by MasaCtrl (Cao et al. 2023) and P2P (Hertz et al. 2022), our proposed ***ClothingNet*** adopts the same UNet of denoising network for appearance extraction. Additionally, we design a ***semantic-refined module*** within the appearance flow to decouple fine-grained features, enhancing the model's understanding of part-level semantics. **Second**, the denoising flow aligns and fuses the structure information from the structure flow with the appearance details from the appearance flow. Rather than directly concatenating or adding external features into the denoising process, as done in methods like ControlNet (Zhang, Rao, and Agrawala 2023), IP2P (Brooks, Holynski, and Efros 2023), and AnyDoor (Chen et al. 2024b), we introduce ***cross-blend attention mechanism*** to capture the relationship between clothing structure and appearance features. **Third**, to evaluate the effectiveness of FashionTailor, we collected a large-scale clothing dataset, ***StructureFashion***, which includes thousands of fashion items with hundreds of combinations of ***neckline types, sleeve types, top lengths, and bottom lengths.*** Ultimately, we created an image-instruction-image triplet dataset of approximately six million scales. Please refer to the **Supplementary Material** for a detailed introduction to the motivation and dataset.

For clarity, the contributions of this paper are summarized as follows:

- We are the first to introduce a clothing structure editing (CSE) task for real human images, posing a challenging part-level editing problem.
- We propose FashionTailor for CSE, a model based on the triple-flow framework. It comprises a ClothingNet-based appearance flow with a semantic-refined module within, a cross-blend attention-based denoising flow to align appearance features and structure information, and an instruction-based structure flow. Finally, FashionTailor supports editing the structure of multiple parts on arbitrary garments simultaneously.
- We created a large-scale clothing structure dataset, StructureFashion, containing about six million pairs. Extensive experiments validated the effectiveness of our model and its zero-shot generalization capability on real wild images.

Related Work

Local Image Editing

Local image editing, as opposed to condition-guided image synthesis, involves modifying an image without changing its primary components. InpaintAnything (Yu et al. 2023) utilizes the object segmentation capabilities of SAM (Kirillov et al. 2023) and the image generation capabilities of Stable Diffusion (Rombach et al. 2022) to replace objects in the source image with textual descriptions. Paint-by-Example (Yang et al. 2023) suggests a method of image editing guided by a reference image. It uses the CLIP (Radford et al. 2021) image encoder to convert the reference image into an embedding, guiding the painting of semantically coherent objects onto a scene image. As a result, this approach may cause a significant loss of detail in the reference image. Other methods (Avrahami, Lischinski, and Fried 2022; Lugmayr et al. 2022) depend on masks to define the areas requiring modification. However, manually obtaining masks in practical applications is labor-intensive, impeding the development of automated intelligent editing (Zou et al. 2024) Recently, methods like P2P (Hertz et al.

2022) and MasaCtrl (Cao et al. 2023) have sparked increased interest in inversion-based local image editing. TIC (Duan et al. 2024), LITS (Jung et al. 2024), FEC (Chen and Huang 2023), and InfEdit (Xu et al. 2023) represent a series of efforts to enhance image consistency and address the quality issues in DDIM inversion reconstruction, employing automatic mask acquisition techniques. DreamMatcher (Nam et al. 2024), LIME (Simsar et al. 2024), and TIGuidedEdit (Wang et al. 2004) further enhance the accuracy of automatic masking. However, these methods require either fine-tuning or time-consuming inversion. Moreover, these methods perform object-level editing operations that preserve the image structure, relying on pre-trained model priors without professional fashion-domain knowledge.

Discussion. To advance image editing techniques in fashion, we design a new part-level editing paradigm suitable for CSE and create a part-level fashion structure dataset.

Fashion Image Editing

The apparel industry plays a vital role in the global economy and stands at the forefront of innovation. In recent years, tasks such as virtual try-on (Du et al. 2024; Shim, Chung, and Heo 2024), outfit generation (Zhang et al. 2024b; Zhou et al. 2023), fashion deconstruction (Li et al. 2024; Yan et al. 2022a), and fashion designing (Baldrati et al. 2023; Yan et al. 2022b) have garnered significant attention. All of these tasks fall into the category of condition-guided fashion image synthesis. In fashion image editing, Kong et al. 2023 demonstrated multi-attribute editing of apparel images using classifier guidance. DiffCloth (Zhang et al. 2023) tackles the alignment between visual and textual representations in apparel design by introducing structural-semantic consensus guidance. Despite these methods enabling preliminary editing of clothing images, they are limited to cabinet clothing and cannot address full-body images with people. Full-body images display clothing folds and their interaction with the wearer, making them more complex and challenging to process but more valuable for practical applications. Moreover, these methods are confined to editing generated clothing images and fail to generalize to real images.

Discussion. In contrast to the existing literature, FashionTailor is the first work to specialize in part-level structure editing of real full-body images.

Methodology

This paper aims to develop an instruction-guided diffusion model for part-level structure editing of human images. Given an original image to be edited and a text-based instruction for structure editing, our model generates a consistency-preserving, high-quality human image, acting like an experienced tailor. The FashionTailor pipeline is demonstrated in Figure 3.

Preliminary

This section begins with a brief introduction of latent diffusion models (LDMs) (Rombach et al. 2022) and IP2P (Brooks, Holynski, and Efros 2023), which compose foundational elements of our methodology. The LDM comprises three main components: a Variational Autoencoder (VAE) (Esser, Rombach, and Ommer 2021) for compressing images from pixel to latent space, a CLIP (Radford et al. 2021) text encoder for processing textual conditions, and a UNet-based network for noise prediction. Further, IP2P extends the generative capabilities of LDMs into intuitive image editing. Mirroring the network construction of the LDM, IP2P integrates image conditions and alters the input prompt conventions. Specifically, given an edited image I_e, the VAE encoder extracts its latent representation z, i.e., $z = \mathcal{E}(I_e)$ that can be reconstructed by the decoder $\mathcal{D}$, i.e., $\tilde{I}_e = \mathcal{D}(z)$. Then, IP2P edits the source image I_s following the given editing instruction T, which the CLIP encoder converts into token embeddings c. During the pre-training stage, noise $\sim \mathcal{N}(0, 1)$ with a level controlled by the timestep t is added to z to create noisy latent z_t. The UNet-based denoiser $\epsilon_\theta(z_t, t, \mathcal{E}(I_s), c)$ is learned to predict the noise and reverse this process. The optimization can be formulated as,

$$\mathcal{L}_{mse} = \mathbb{E}_{z,\epsilon,\mathcal{E}(I_s),c,t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \mathcal{E}(I_s), c)\|\right]. \quad (1)$$

To allow unconditional denoising, two conditions are randomly omitted during training with a certain probability by setting $I_s = \emptyset_I$ or $c = \emptyset_c$. During the sampling process, the score estimate is as follows:

$$\begin{aligned}\tilde{\epsilon}_\theta(z_t, t, I_s, c) &= \epsilon_\theta(z_t, t, \emptyset_I, \emptyset_c) \\ &+ S_I(\epsilon_\theta(z_t, t, I_s, \emptyset_c) - \epsilon_\theta(z_t, t, \emptyset_I, \emptyset_c)), \\ &+ S_c(\epsilon_\theta(z_t, t, I_s, c) - \epsilon_\theta(z_t, t, I_s, \emptyset_c))\end{aligned} \quad (2)$$

two guidance scales, S_I and Sc, are introduced. Increasing S_I brings the edited image closer to the source image while increasing S_c yields more intense editing.

FashionTailor

As illustrated in Figure 3, we present the overall pipeline of our method, which consists of three basic interacting workflows: (1) Appearance flow, which captures appearance details at various scales, such as clothing texture and human features. Our proposed ClothingNet extracts both low-level features from the VAE encoder and high-level decoupled features enhanced by a semantic-refined module, improving the model's understanding of the relationship between humans and clothing; (2) Structure flow, which encodes instructions for controllable editing of the clothing structure in the source image; (3) Denoising flow, which denoises the latent image while effectively integrating the two conditions. Unlike the replacement mechanisms of training-free methods such as MasaCtrl (Cao et al. 2023) and P2P (Hertz et al. 2022), we propose a cross-blend attention mechanism. It explores various effective fusion methods to ensure alignment of appearance details with clothing structure and maintain consistency between the human body and clothing features after editing.

Detail Feature Extraction

The CSE task involves non-rigid editing, necessitating consistent visual content generation for deformed regions. This

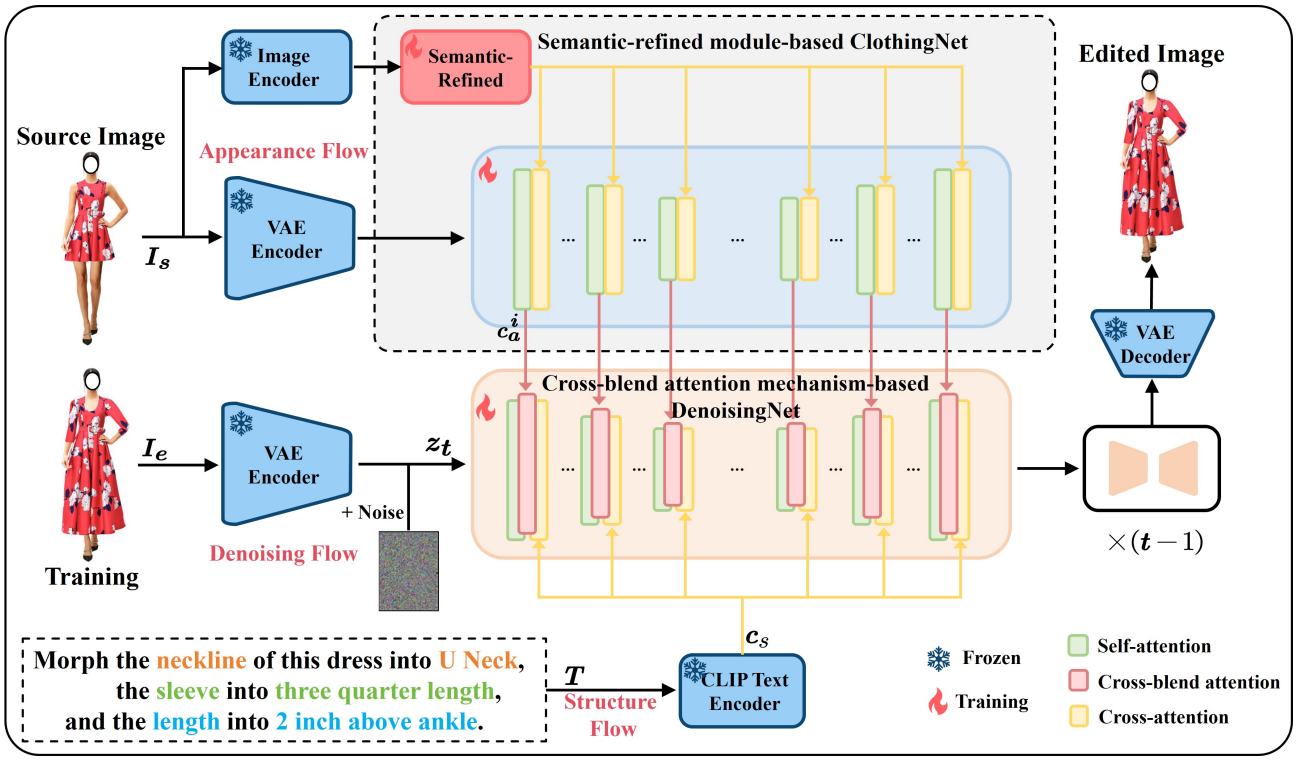

Figure 3: **An overview of the proposed FashionTailor framework.** Our model consists of (1) Appearance flow, which encodes the appearance features of the source image with a semantic-refined module for refining the semantic information; (2) Structure flow, which processes the target instruction; and (3) Denoising flow, which precisely aligns target structural information and appearance features based on the cross-blend attention mechanism.

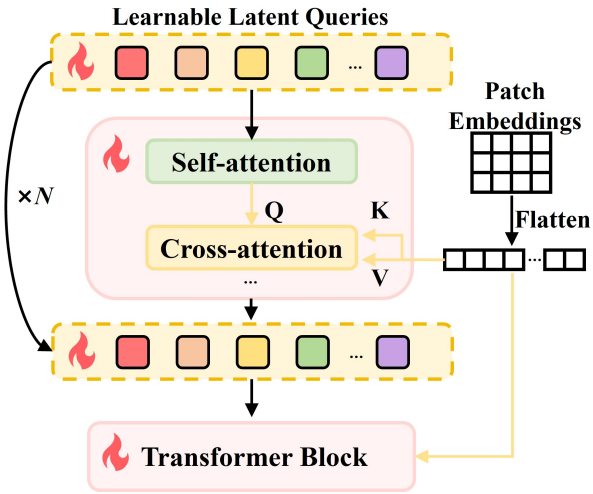

Figure 4: **Semantic-refined module.** Patch embeddings interact via cross-attention at each transformer block.

may include the generation of both human skin and additional clothing textures. Previous methods (Yu et al. 2023) often used the CLIP image encoder as the feature extractor. Since CLIP mainly focuses on aligning with high-level textual features, it lacks fine-grained encoding of appearance features. To overcome this, we introduce ClothingNet as the appearance extractor. It shares the same architecture as the denoising network, ensuring that features at each scale reside in the same feature space. Unlike the denoising UNet, ClothingNet does not involve a diffusion process of adding noise to the source image, as it is solely used for feature extraction.

To enable ClothingNet to focus on areas relevant to the editing instruction adaptively, we design a semantic-refined module after the image encoder, as shown in Figure 4. The image encoder can be any model that extracts semantic-level information, such as the CLIP image encoder or ViT (Liu et al. 2021). In addition, the initial flattened image embedding from the encoder is a global, coarse-grained representation that lacks understanding of the parts of the human image. Apart from altering clothing structure, the CSE process needs to effectively characterize the complex structure of human images consisting of body parts, clothing parts, and possibly intricate backgrounds. As a result, it is essential to decouple the desired part-level semantics. To achieve this, we maintain a set of learnable queries representing the different semantics of human images with the usage of standard transformer blocks. The patch embeddings output by the image encoder serve as the input for the semantic-refined module, interacting via the cross-attention module at each transformer block.

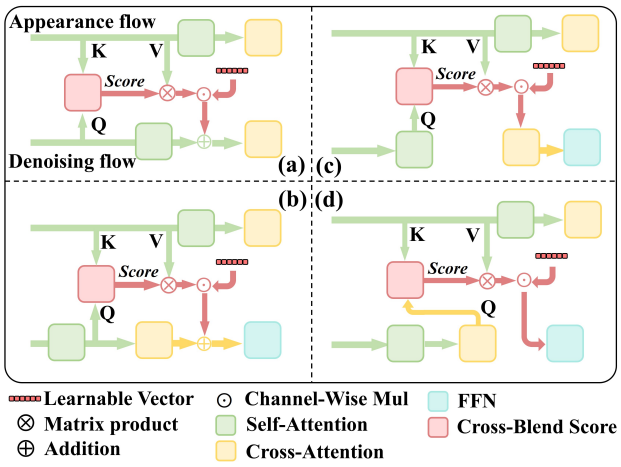

Figure 5: **Cross-blend attention mechanism:** (a) self- and (b) cross-enhanced blend attention, (c) image-text and (d) text-image dual-cross attention.

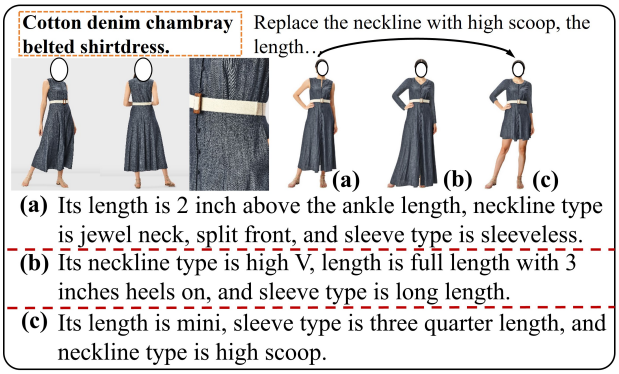

Figure 6: A sampled product from our StructureFashion.

Cross-blend Attention

Global shape control features c_s from the structure flow provide identical conditions for blocks at each scale of denoising UNet. Meanwhile, multi-scale appearance features c_a from the appearance flow provide as many accurate details as possible. To align c_s and c_a, as well as precise control of the texture details and structure of generated images, we introduce a cross-blend attention (CBA) mechanism. As shown in Figure 5, we design four attention operators that exploit effective feature fusion methods within the CBA mechanism. Specifically, the self- and cross-attention layers of the diffusion model contain extensive structure and appearance information for image generation. The image layout can be roughly formed in queries, covering the semantic changes complying with the target prompt. The keys and values of self-attention represent the content features during the synthesis process (Cao et al. 2023; Wang, Liu, and Xu 2024).

Therefore, the query of the CBA should derive from the denoising flow, which incorporates the target structure information, while the key and value should stem from the self-attention layer of ClothingNet. Based on the different options to query and fusion feature injection, we design the following strategies: self-enhanced blend attention, where the query originates from the self-attention in the denoising UNet, and the fusion feature blends with the self-attention output (Figure 5a); cross-enhanced blend attention, where the query originates from the cross-attention in the denoising UNet, and the fusion feature blends with the cross-attention output (Figure 5b). Both strategies enhance feature integration and are defined as follows:

$$F_o = \text{softmax}\left(\frac{Q(K)^\top}{\sqrt{d}}\right)V + \lambda \odot \text{softmax}\left(\frac{Q(K\prime)^\top}{\sqrt{d}}\right)V\prime \quad (3)$$

where Q, K, and V derive from the self- or cross-attention of denoising UNet, $K\prime$, $V\prime$ are from the corre-sponding self-attention layer of ClothingNet, and λ is the learnable fusion vector.

Additionally, since source and edited images are essentially the same objects with different structures in the CSE task, source images can be regarded as a different modality of edited images (Zhang et al. 2024b). The instruction with text modality is semantically aligned with the edited image, while the source image with image modality is aligned with the edited image in terms of texture details. Consequently, in line with the general design of the attention mechanism, we incorporate an additional cross-attention layer after the self-attention to align the texture information of the image modality. Specifically, we design image-text dual-cross attention and text-image dual-cross attention, as illustrated in Figures 5(c) and 5(d), respectively. The primary difference between these two lies in the order of the cross-attention layers. In the experimental section, we have fully evaluated the effectiveness of these designs. Notably, the loss function and classifier-free guidance of our FashionTailor adhere to equations (1) and (2), respectively.

Experiments

Implementation Details

Dataset. Currently, available datasets for fashion images are for person image generation and virtual try-on, lacking the required part-level multimodal pairwise information needed to perform the CSE task. To advance image editing techniques in fashion, we collected the first clothing structure dataset, as shown in Figure 6. Please refer to the **Supplementary Material** for more details.

Hyperparameters. In this study, we utilized pretrained IP2P (Brooks, Holynski, and Efros 2023) and Stable Diffusion (Rombach et al. 2022) v1.5 to initialize the weights of the denoising UNet and ClothingNet, respectively. For classifier-free guidance, we set S_I and S_c to 1.5 and 7.5, respectively, and drop the conditions I_s and c with a probability of 5%. Due to computational resource limitations, we did not use all available data for model training. Please refer to the **Supplementary Material** for more details.

Baselines. We compare FashionTailor with the state-of-the-art text-to-image methods based on reference images,

Category	Method	Publication	FID↓	KID↓	LPIPS↓	SSIM↑	PSNR↑	CLIP-S↑	D-CLIP↑
Inversion	MasaCtrl	ICCV'23	76.77	4.723	0.267	0.700	15.859	8.958	0.172
	NTI+FPE	CVPR'24	36.55	1.228	0.232	0.723	16.508	11.412	0.167
	InfEdit	CVPR'24	31.69	1.237	0.171	0.782	17.118	17.631	0.092
Adapter	IP-Adapter	arXiv'23	25.62	0.758	0.316	0.534	13.210	12.423	0.252
	MagicClothing	ACMMM'24	67.81	4.154	0.253	0.745	16.581	17.769	0.287
Instruction	MagicBrush	NeurIPS'24	27.41	0.840	0.220	0.716	16.780	8.680	0.179
	IP2P	CVPR'23	25.37	0.600	0.057	0.904	23.761	37.411	0.699
	FashionTailor	This work	**24.70**	**0.515**	**0.048**	**0.916**	**24.444**	**38.642**	**0.731**

Table 1: Quantitative comparison of our FashionTailor with the state-of-the-art text-to-image methods based on references.

Method	KID↓	LPIPS↓	SSIM↑	CLIP-S↑	D-CLIP↑
B1	0.600	0.569	0.904	37.411	0.699
B2	0.555	0.515	0.912	38.451	0.726
B3	0.530	0.512	0.913	38.520	0.725
B4	0.526	0.510	0.913	38.661	0.730
B5	0.534	0.542	0.910	**38.672**	0.726
B6	0.523	0.509	0.913	38.526	0.727
B7	**0.515**	**0.486**	**0.916**	38.642	**0.731**

Table 2: Quantitative results for ablation studies.

including the instruction-guided IP2P (Brooks, Holynski, and Efros 2023) and MagicBrush (Zhang et al. 2024a), the adapter-based IP-Adapter (Ye et al. 2023) and MagicClothing (Chen et al. 2024a), as well as the inversion-based MasaCtrl (Cao et al. 2023), NTI+FPE (Liu et al. 2024), and InfEdit (Xu et al. 2024). For a fair comparison, we fine-tuned IP2P, IP-Adapter and MagicClothing on StructureFashion. Additionally, source and target descriptions were provided for inversion-based methods.

Comparisons

Quantitative Results. As reported in Table 1, our method significantly outperforms the state-of-the-art methods across all metrics. (1) Notably, despite the extensive research on inversion-based image editing, these methods are unsuitable for the CSE task. On one hand, these methods rely on the priors of pre-trained models that lack fashion expertise. On the other hand, they prefer object-level edits with more obvious semantics and struggle to capture part-level concepts. (2) For adapters, we conduct fine-tuning on StructureFashion as reported in their original papers. IP-Adapter achieves relatively better FID and KID scores but performs poorly on consistency scores. This indicates that although it can generate decent-quality images, it deviates from the editing trajectory, shows significant over-editing, and struggles to maintain consistency in irrelevant regions. In contrast, MagicClothing achieves better consistency scores but performs poorly in fidelity generation. These results for both categories of methods indicate that they are unsuitable for the CSE task due to its complexity. (3) IP2P shows suboptimal results across all metrics, demonstrating its inability to capture fine-grained features accurately. (4) Compared to these methods, our FashionTailor, benefiting from the design of appearance flow and cross-blend attention, exhibits superior results in terms of the balance of fidelity generation, editability, and consistency preservation.

Qualitative Results. Figure 7 illustrates a comprehensive comparison of qualitative results with existing alternatives for the CSE task. (1) As we can see, inversion-based methods cannot follow the editing instructions, and the clothing structure of the generated image remains unchanged. Additionally, these methods fail to accurately reconstruct the complex clothing texture information, as shown in columns 3-5. (2) IP-Adapter weakly follows the instruction but struggles to maintain clothing details and character identity, often generating a variant of the source image. Despite generating instruction-consistent clothing structures, MagicClothing performs the worst in clothing detail generation, possibly due to the framework's incompatibility with the CSE task. (3) IP2P shows visual results similar to ours but fails to reasonably predict the new clothing textures. Specifically, IP2P tends to lose the edge texture features of the clothing, as shown in the highlighted sections of columns 9-10. However, as humans are sensitive to subtle changes in garment features (e.g., patterns), proper editing of clothing structure is not a simple matter of cutting or adding fabric. It requires complete consistency of the clothing. (4) Compared to these methods, FashionTailor produces the most outstanding results, strictly following the instructions and faithfully maintaining the consistency of clothing and human body features.

Ablation Study

To verify the effectiveness of the appearance flow and cross-blend attention mechanisms, we explore various alternative designs, as reported in Table 2. (1) B1 represents the configuration without appearance flow and cross-blend attention, where the source image is directly added to the UNet features at each scale after passing through the VAE encoder. (2) We then improve the network by introducing ClothingNet in B2 to achieve multi-scale fine-grained feature injection. In this approach, we directly concatenate the outputs of each attention module in ClothingNet with the inputs of the corresponding attention module in the denoising UNet. From Table 2, it is evident that the introduction of ClothingNet significantly improves the generation quality. (3) Furthermore, we design cross-blend attention to enhance

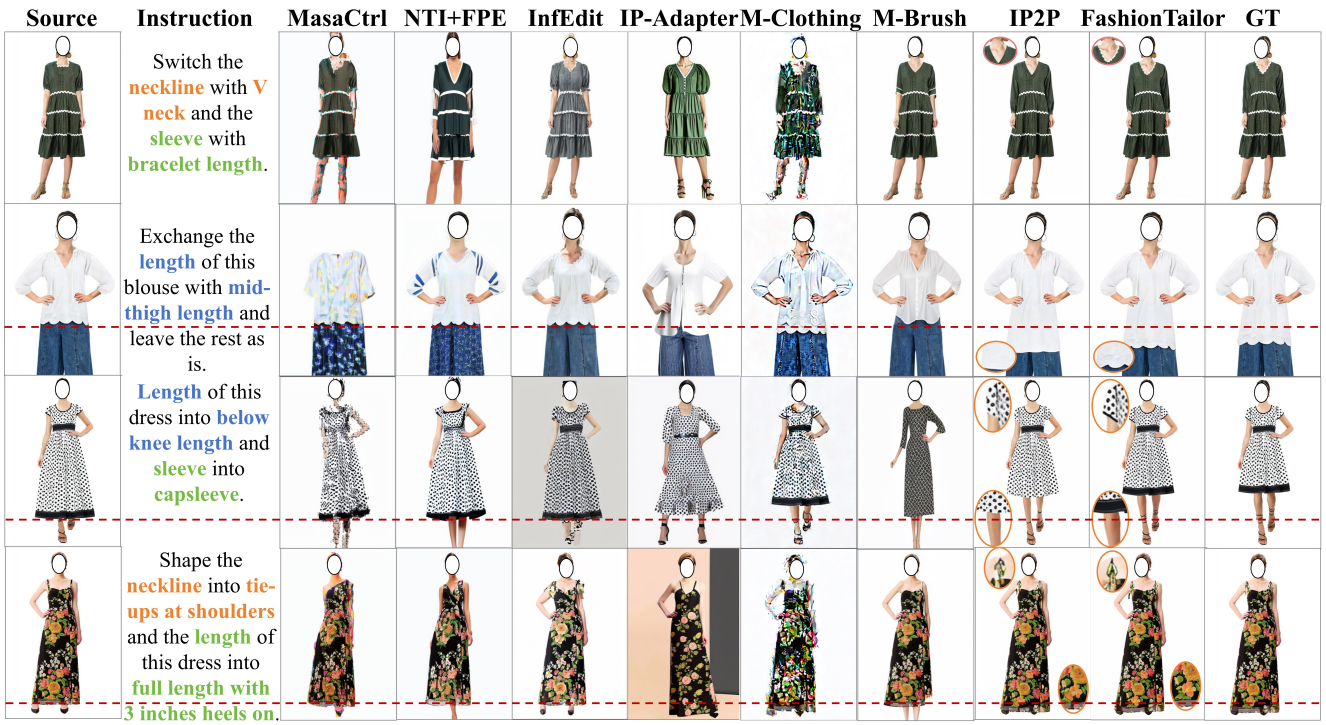

Figure 7: Qualitative comparisons with existing alternatives for CSE, including MasaCtrl, NTI+FPE, InfEdit, IP-Adapter, MagicClothing, MagicBrush, and IP2P. The red dotted line makes it easier to see the change in length.

Figure 8: Examples of our FashionTailor applications in real scenarios.

the injection of fine-grained features. As shown in Figure 5, we design four blending modes: B3-B6 denote image-text, text-image dual-cross and cross-, self-enhanced blend attention, respectively. We observe a significant improvement in fidelity and CLIP scores for these four attention modes. Although cross-enhanced blend attention achieves the highest CLIP scores, it performs the worst in KID, LPIPS, and SSIM. As CLIP scores focus on semantic relations and overlook detail matching, this design may cause over-editing. In addition, self-enhanced blend attention chosen as the default configuration for the subsequent experiments achieves the best generation quality. (4) Moreover, cross-blend attention only slightly improves the image generation quality. We believe that the indiscriminate injection of redundant fine-grained information during the denoising process is the root cause. To address this, we designed the semantic-refined module in B7 to assist ClothingNet in concentrating on regions that need to be modified. Compared to B6, all metrics show significant improvement, highlighting the effectiveness of the semantic-refined module.

Real Scenario Application. Figure 8 illustrates the generalization capability of FashionTailor in real-world scenarios, showing that the model learns to interact effectively with clothing structures and texture details without overfitting.

Conclusion

We present a new framework for part-level image editing called clothing structure editing. Our proposed FashionTailor utilizes a triple-flow method to balance editability and consistency preservation through a cross-blend attention mechanism and a semantic-refined module. Additionally, we constructed StructureFashion, a dataset comprising six million image-instruction-image triples. Our approach not only surpasses existing methods but also demonstrates promising results in real-world applications.

Acknowledgments

This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant no. 2021B1515020088, and the Shenzhen Science and Technology Program under Grant no. JCYJ20240813104843058. Furthermore, the work was also supported by the E-Business & Enterprise Intelligent Computing Research Center, Big Data Research Center of Harbin Institute of Technology, Shenzhen.

References

Avrahami, O.; Lischinski, D.; and Fried, O. 2022. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18208–18218.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23393–23402.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18392–18402.

Cao, M.; Wang, X.; Qi, Z.; Shan, Y.; Qie, X.; and Zheng, Y. 2023. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22560–22570.

Chen, C. 2020. Analysis of the Influence of Fashion Design on Clothing Structure and Garment Management. In *Recent Trends in Decision Science and Management: Proceedings of ICDSM 2019*, 39–44. Springer.

Chen, S.; and Huang, J. 2023. Fec: Three finetuning-free methods to enhance consistency for real image editing. In *2023 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*, 76–87. IEEE.

Chen, W.; Gu, T.; Xu, Y.; and Chen, A. 2024a. Magic clothing: Controllable garment-driven image synthesis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 6939–6948.

Chen, X.; Huang, L.; Liu, Y.; Shen, Y.; Zhao, D.; and Zhao, H. 2024b. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6593–6602.

Du, C.; Wang, J.; Rong, Y.; Liu, S.; Liu, K.; and Xiong, S. 2024. CycleVTON: A Cycle Mapping Framework for Parser-Free Virtual Try-On. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1618–1625.

Duan, X.; Cui, S.; Kang, G.; Zhang, B.; Fei, Z.; Fan, M.; and Huang, J. 2024. Tuning-free inversion-enhanced control for consistent image editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1644–1652.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12873–12883.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. C. 2017. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-Prompt Image Editing with Cross Attention Control. arXiv:2208.01626.

Hu, Y. 2023. Analysis of clothing structure and management in clothing design oriented to market demand via recommendation algorithm. *Electronic Commerce Research*, 1–22.

Jung, Y.; Lee, S.; Djanibekov, T.; Shim, H.; and Ye, J. C. 2024. Latent Inversion with Timestep-aware Sampling for Training-free Non-rigid Editing. arXiv:2402.08601.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Kong, C.; Jeon, D.; Kwon, O.; and Kwak, N. 2023. Leveraging off-the-shelf diffusion model for multi-attribute fashion image manipulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 848–857.

Li, N.; Liu, Q.; Singh, K. K.; Wang, Y.; Zhang, J.; Plummer, B. A.; and Lin, Z. 2024. UniHuman: A Unified Model For Editing Human Images in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2039–2048.

Liu, B.; Wang, C.; Cao, T.; Jia, K.; and Huang, J. 2024. Towards Understanding Cross and Self-Attention in Stable Diffusion for Text-Guided Image Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7817–7826.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11461–11471.

Miyato, T.; Kataoka, T.; Koyama, M.; and Yoshida, Y. 2018. Spectral Normalization for Generative Adversarial Networks. arXiv:1802.05957.

Nam, J.; Kim, H.; Lee, D.; Jin, S.; Kim, S.; and Chang, S. 2024. Dreammatcher: Appearance matching self-attention for semantically-consistent text-to-image personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8100–8110.

Nobile, T. H.; and Cantoni, L. 2023. Personalization and customization in fashion: searching for a definition. *Journal of Fashion Marketing and Management: An International Journal*, 27(4): 665–682.

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*, 8748–8763. PMLR.

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*, 10684–10695.

Shim, S.-H.; Chung, J.; and Heo, J.-P. 2024. Towards squeezing-averse virtual try-on via sequential deformation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4856–4863.

Simsar, E.; Tonioni, A.; Xian, Y.; Hofmann, T.; and Tombari, F. 2024. LIME: Localized Image Editing via Attention Regularization in Diffusion Models. arXiv:2312.09256.

Wang, J.; Liu, P.; and Xu, W. 2024. Unified Diffusion-Based Rigid and Non-Rigid Editing with Text and Image Guidance. arXiv:2401.02126.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Xie, Z.; Huang, Z.; Zhao, F.; Dong, H.; Kampffmeyer, M.; Dong, X.; Zhu, F.; and Liang, X. 2022. PASTA-GAN++: A Versatile Framework for High-Resolution Unpaired Virtual Try-on. arXiv:2207.13475.

Xu, S.; Huang, Y.; Pan, J.; Ma, Z.; and Chai, J. 2023. Inversion-Free Image Editing with Natural Language. arXiv:2312.04965.

Xu, S.; Huang, Y.; Pan, J.; Ma, Z.; and Chai, J. 2024. Inversion-Free Image Editing with Language-Guided Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9452–9461.

Yan, H.; Zhang, H.; Liu, L.; Zhou, D.; Xu, X.; Zhang, Z.; and Yan, S. 2022a. Toward intelligent design: An ai-based fashion designer using generative adversarial networks aided by sketch and rendering generators. *IEEE Transactions on Multimedia*, 25: 2323–2338.

Yan, H.; Zhang, H.; Shi, J.; and Ma, J. 2022b. Texture brush for fashion inspiration transfer: A generative adversarial network with heatmap-guided semantic disentanglement. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(5): 2381–2395.

Yang, B.; Gu, S.; Zhang, B.; Zhang, T.; Chen, X.; Sun, X.; Chen, D.; and Wen, F. 2023. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18381–18391.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. arXiv:2308.06721.

Yu, T.; Feng, R.; Feng, R.; Liu, J.; Jin, X.; Zeng, W.; and Chen, Z. 2023. Inpaint Anything: Segment Anything Meets Image Inpainting. arXiv:2304.06790.

Zhang, J.; Li, K.; Lai, Y.-K.; and Yang, J. 2021. Pise: Person image synthesis and editing with decoupled gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7982–7990.

Zhang, K.; Mo, L.; Chen, W.; Sun, H.; and Su, Y. 2024a. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, X.; Lin, E.; Li, X.; Luo, Y.; Kampffmeyer, M.; Dong, X.; and Liang, X. 2024b. MMTryon: Multi-Modal Multi-Reference Control for High-Quality Fashion Generation. arXiv:2405.00448.

Zhang, X.; Yang, B.; Kampffmeyer, M. C.; Zhang, W.; Zhang, S.; Lu, G.; Lin, L.; Xu, H.; and Liang, X. 2023. Diffcloth: Diffusion based garment synthesis and manipulation via structural cross-modal semantic alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23154–23163.

Zhou, D.; Zhang, H.; Ma, J.; Fan, J.; and Zhang, Z. 2023. Fcboost-net: A generative network for synthesizing multiple collocated outfits via fashion compatibility boosting. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7881–7889.

Zou, S.; Tang, J.; Zhou, Y.; He, J.; Zhao, C.; Zhang, R.; Hu, Z.; and Sun, X. 2024. Towards Efficient Diffusion-Based Image Editing with Instant Attention Masks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7864–7872.