PICTURE: PhotorealistIC virtual Try-on from UnconstRained dEsigns

Shuliang Ning[1,2,3] Duomin Wang[3] Yipeng Qin[4] Zirong Jin[2]
Baoyuan Wang[3] Xiaoguang Han[2,1,*]
[1] FNii, CUHKSZ [2] SSE, CUHKSZ [3] Xiaobing.AI [4] Cardiff University

Figure 1. Examples of virtual try-on manipulations using ucVTON on **in-the-wild** real images. Orange box: style control; Green box: texture control; Blue box: design elements control.

Abstract

In this paper, we propose a novel virtual try-on from unconstrained designs (ucVTON) task to enable photorealistic synthesis of personalized composite clothing on input human images. Unlike prior arts constrained by specific input types, our method allows flexible specification of style (text or image) and texture (full garment, cropped sections, or texture patches) conditions. To address the entanglement challenge when using full garment images as conditions, we develop a two-stage pipeline with explicit disentanglement of style and texture. In the first stage, we generate a human parsing map reflecting the desired style conditioned on the input. In the second stage, we composite textures onto the parsing map areas based on the texture input. To represent complex and non-stationary textures that have never been achieved in previous fashion editing works, we first propose extracting hierarchical and balanced CLIP features and applying position encoding in VTON. Experiments demonstrate superior synthesis quality and personalization enabled by our method. The flexible control over style and texture mixing brings virtual try-on to a new level of user experience for online shopping and fashion design.

* Corresponding author: hanxiaoguang@cuhk.edu.cn.

1. Introduction

Virtual try-on (VTON) systems have become indispensable in the era of online clothing shopping and are an active area of research in computer vision. As online shopping grows increasingly ubiquitous in modern digital lifestyles, VTON addresses a key limitation: the inability to physically try on clothes prior to purchase. By enabling users to visualize garments on their photos or avatars, VTON systems aim to provide crucial visual information about fit and appearance.

Traditional VTON [2, 6, 7, 9, 11, 12, 18, 21, 22, 38, 39] focuses on in-shop scenarios which create photorealistic visualization by generating images of individuals wearing existing retail garments. This is constrained by input person and clothing images so that the person's identity and appearance are retained and the generated clothing matches the input. However, this constrained generation is also limited, as real-world online shopping behaviors often involve a bit of design, with a desire to mix-and-match elements from different garments. That is, users may wish to visualize personalized composite clothing items, combining preferred style and texture aspects from separate pieces. Enabling such controllable synthesis of new clothing on virtual try-on remains an open challenge. FashionTex [19] pioneered this strand by using a disentangled (style, texture) representation to control the VTON synthesis. However, despite its success, FashionTex is still constrained by the type of its inputs: i) the style condition is limited to text prompts, which are less accessible and inclusive to users with dyslexia or other language barriers; ii) the texture input is limited to small image patches, which cannot characterize complex fabrics or patterns. Both constraints reduce its real-world applicability.

In this work, we fill this gap by proposing a new task called VTON from unconstrained designs (ucVTON), which greatly relaxes the constraints mentioned above, enabling users to specify style via images, and texture via full-garment images or swatches. Specifically, we expand the allowable input types for increased flexibility: the style condition can be a text prompt or an example garment image; the texture condition can be a full garment, a cropped section of a garment, or an image patch. This allows representing complex fabric textures across different scales and spatial distributions, bringing fine-grained control of VTON to a new level of synthesis quality and user personalization. However, a significant challenge emerges when enabling style and texture inputs to be full garment images: both inputs contain entangled style and texture features, and the network can struggle to disentangle the irrelevant style from the texture input and irrelevant texture from the style input for proper re-composition.

Addressing this challenge, we propose a novel two-stage style and texture disentanglement pipeline based on Stable Diffusion [26], where the first stage explicitly learns to generate clothing styles in the format of a human parsing map conditioned on a given style input; the second stage then adds textures to the garment areas of the parsing map conditioned on the given texture input. In addition, instead of naively replacing the text CLIP features used in previous methods with image CLIP features, we propose a novel approach based on our observation that the final-block CLIP features are rich in semantics but lack low-level texture features. Specifically, we propose to learn photorealistic texture features from the texture input through a hierarchical and balanced combination of CLIP features from multiple blocks of its Vision Transformer. Finally, we utilize position encoding [30] to represent the spatial distributions and scales of non-stationary input textures. This significantly expands the space of possible texture inputs that can be handled, which has never been achieved by previous methods. Extensive experimental results demonstrate the effectiveness of our method. Our contributions include:

- We introduce the novel task of virtual try-on from unconstrained designs (ucVTON), which significantly advances the state-of-the-art by enabling photorealistic VTON from diverse style (text prompts or garment images) and texture inputs (full garments, cropped sections of a garment, or image patches).
- We propose a two-stage disentanglement pipeline that explicitly separates style generation and texture composition for controllable ucVTON from entangled inputs.
- We first propose a new hierarchical CLIP feature extraction and position encoding method to represent photorealistic, non-stationary textures. This significantly expands the diversity and complexity of synthesized textures that has never been achieved by previous methods.
- Extensive experiments demonstrate that our method has brought fine-grained control of VTON to a new level of synthesis quality and user personalization.

2. Related Work

Image-based Fashion Editing. Given a human image and some editing conditions, such as reference cloth images [5, 11, 14, 22, 37] or text descriptions of styles [1, 3, 15, 24, 34], image-based fashion editing aims to generate a target image that satisfies the given conditions while preserving the rest (*e.g.*, identity, skin) of the original human image. Examples of existing solutions include TextureReformer [31], which employs a multi-view, multi-stage synthesis procedure to perform interactive texture transfer under user-specified guidance; Text2Human [15], which implements a two-step process using text descriptions to synthesize human images from a given pose, focusing on clothing shapes and textures; FashionTex [19], which develops a fashion editing module that harnesses both text and texture inputs for multi-level fashion editing in full-body portraits.

In this work, we follow the multimodal setup of FashionTex [19] and take both text and example images as inputs. However, unlike [19] which assumes the input texture to be a small image patch (64×64) with stationary textures, we have made significant improvements to the model to enable it to handle unconstrained high-resolution garment images with arbitrary textures, shapes, positions, and scales as input. This allows for higher quality, flexibility and control over the generated results compared to prior arts.

Diffusion Models. Diffusion models [13, 26, 28] are a major type of deep generative models that exhibit robustness and superior proficiency in handling diverse data modalities [41]. Inspired by [23, 26, 40], numerous conditional image generation works [10, 17, 25, 27, 32, 35, 36] have incorporated diffusion models into their methods. For example, SGDiff [29] introduces a style-guided diffusion model that combines text and style images to facilitate creative image synthesis; Paint-by-Example [37] leverages self-supervised training to disentangle and re-organize the source image and the exemplar for more precise control in exemplar-guided image inpainting; Multimodal Garment Designer [3] employs a latent diffusion model and multimodal conditions such as text, pose map, and garment sketches for fashion editing. Their impressive results demonstrate the power of diffusion models as a solution for high-quality, controllable image generation and editing tasks.

Method	C_s^{text}	C_s^{img}	C_t^{patch}	C_t^{img}
Texture Reformer[31]	✗	✗	✓	✓
PIDM[4]	✗	✗	✗	✓
FashionTex[19]	✓	✗	✓	✗
Text2Human[15]	✓	✗	✗	✗
DCI-VTON[11]	✗	✓	✓	✗
Ours	✓	✓	✓	✓

Table 1. Comparison with SOTA methods on input types allowed.

3. Background and Problem Definition

Traditionally, virtual try-on can be defined as:

$$\hat{I} = \text{VTON}(I, G) \tag{1}$$

where I denotes the input image of a person, G denotes the garment image to be virtually worn, $\hat{I}$ is the output image showing the person wearing garment G. Although straightforward, this definition is limited because it requires a *complete* garment image G with a fixed combination of clothing style and texture. This contradicts people's imagination ability to envision wearing a particular clothing style with different textures. Addressing this limitation, FashionTex [19] pioneered in disentangling G into style condition C_s^{text} and texture condition C_t^{patch} and have:

$$\hat{I} = \text{VTON}(I, \{C_s^{text}, C_t^{patch}\}) \tag{2}$$

where C_s^{text} is a text prompt describing the garment style and C_t^{patch} is a small image patch describing the local texture of the garment. Despite its success, FashionTex is constrained by the type of inputs in that: i) the style condition is limited to text prompts C_s^{text}, which are less accessible and inclusive to users with dyslexia or other language barriers; ii) the texture input is limited to small image patches C_t^{patch}, which cannot characterize complex fabrics or patterns. These constraints reduce its real-world applicability. For practical virtual try-on, users need more flexible condition types - the ability to specify style via images , and texture via full-garment images or swatches. To fill this gap, in this work, we propose virtual try-on from **unconstrained designs** (ucVTON):

$$\hat{I} = \text{ucVTON}(I, \{C_s^{text/img}, C_t^{img/patch}\}) \tag{3}$$

In this way, we expand the allowable condition types for style and texture as follows: the style condition can be either a text prompt or an example garment image, providing more flexibility in specifying the desired design. The texture condition can be a full-garment image, a cropped section of a garment, or an image patch, enabling the representation of complex textures at varied scales and distributions. This enhanced flexibility empowers users with finer-grained control over both style and texture. Please see Table 1 for a comparison of our ucVTON and SOTA methods.

4. Method

As mentioned above, to enable virtual try-on from unconstrained designs, we propose a novel two-stage pipeline based on Stable Diffusion [26] to disentangle clothing types (*e.g.*, long sleeves) and textures from full-body human image data (Sec. 4.1). In addition, to further improve the texture quality, we propose to learn photorealistic texture features in a hierarchical and balanced way (Sec. 4.2). Finally, to endow the model with the ability to learn non-stationary textures as well as the scale and position of textures on the input garment, we propose to incorporate a positional encoding module in our pipeline (Sec. 4.3). An overview of our pipeline is shown in Fig. 2.

Definitions of Additional Symbols. In addition to the symbols defined in Sec. 3, we have:

- $\mathbf{I^m}$: the clothing-agnostic version of I, which is generated by i) masking the whole bounding box area of a garment in I and ii) copy-pasting the hair area of I back on top. These can be easily achieved using the parsing map of I. Note that for lower garments, we ensure the bottom edges of their bounding boxes always go down to the shoes to avoid the leakage of garment length (*e.g.*, short skirts).
- $\mathbf{P(I)}$, $\mathbf{P(I^m)}$: the human parsing images of I and I^m.
- $\mathbf{D(I)}$: the densepose image of I generated by applying readily available methods [33] on the SHHQ dataset [8].
- $\mathbf{P(I^m, C_s)}$: the output parsing map generated by inpainting $P(I^m)$ with the guidance of C_s^{text} or C_s^{img}.
- $\mathbf{z_t^P}$, $\mathbf{z_t^i}$: the latent variable of output parsing $P(I^m, C_s)$ and output human image $\hat{I}$.

4.1. Two-Stage Style and Texture Disentanglement

To make virtual try-on effective with unconstrained designs, we need to disentangle their style and texture so that the style of the texture-conditioning image C_t^{img} (*e.g.*, trousers) and the texture of the style conditions C_s^{img}/ C_s^{text} do not affect the results. This is a challenging task for Stable Diffusion [26], the foundation model we use, as the embeddings generated by its CLIP encoder contain both style and texture information of the input. To address this issue, we propose a novel two-stage pipeline for style and texture disentanglement, where the first stage aims to generate a parsing map whose clothing style is determined by C_s^{img} or C_s^{text}, and the second stage adds clothing texture conditioned on C_t^{img} to the garment areas of the parsing map. Note that we only used C_t^{img} during training and the learned disentanglement allows for using C_t^{patch} during inference.

Stage 1: Parsing-based Style Editing. Given the masked clothing-agnostic human parsing image $P(I^m)$ and its corresponding densepose image $D(I)$ where the arm or leg positions of I are provided, we aim to generate an output parsing map $P(I^m, C_s)$ in which the relevant garment region of

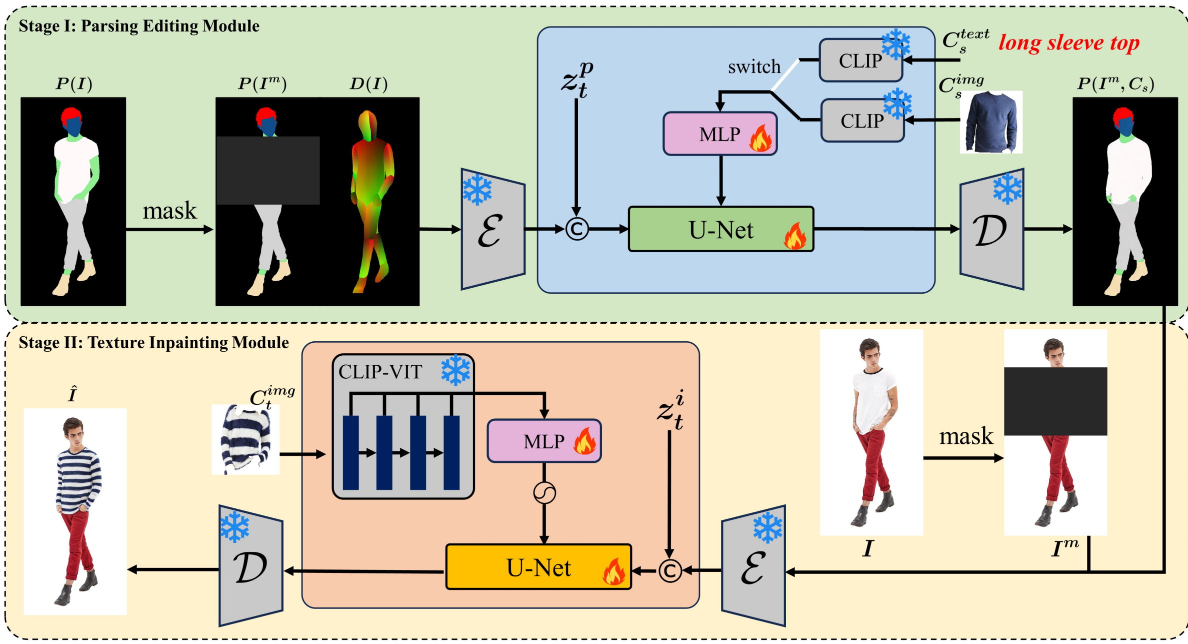

Figure 2. Overall pipeline. **Stage I**: Given a clothing-agnostic parsing image $P(I^m)$ and its corresponding densepose image $D(I)$, a style condition C_s (text or image), this stage generates a parsing map $P(I_m, C_s)$ edited according to C_s. **Stage II**: Given a clothing-agnostic human image I^m, the parsing map $P(I_m, C_s)$ generated in Stage I, a garment texture condition C_t (image or patch), this stage generates the final human image $\hat{I}$ with its style specified by C_s and texture specified by C_t. Note that we only used C_t^{img} during training and the learned disentanglement allows for using C_t^{patch} during inference. Flame symbol: "trainable"; Snowflake symbol: "freeze".

$P(I^m)$ is edited according to C_s^{img} or C_s^{text}. Specifically, we freeze the encoder $\mathcal{E}$, decoder $\mathcal{D}$ and the CLIP encoder of Stable Diffusion and finetune its U-Net component using the inputs mentioned above. Note that we included an MLP layer between the CLIP encoder and U-Net as an adaptor to map the CLIP features to the garment style space. Following [26], we pass $P(I^m)$ and $D(I)$ through the encoder $\mathcal{E}$ and concatenate their embeddings with latent variable z_t^p along the channel dimension as input to the U-Net and have:

$$\epsilon_t = \epsilon_\theta([\mathcal{E}(D(I)), \mathcal{E}(P(I^m)), z_t], t, C_s) \tag{4}$$

where $C_s = \text{MLP}([C_s^{text}, C_s^{img}])$, the sizes of C_s^{text} and C_s^{img} are 1×768. During training, either C_s^{text} or C_s^{img} is set to 0 to ensure that only one style signal is used as input.

Remark. Our parsing-based style editing module inherently enables sequential editing by simply replacing $P(I)$ with the generated $P(I^m, C_s)$ and running the module again. Additionally, in the special case when the source human is wearing separate top and bottom garments, and the target garment style is a dress or jumpsuit, the mask should cover all garment parts to implement the editing properly.

Stage 2: Style-guided Garment Texture Inpainting. Using the edited human parsing map $P(I^m, C_s)$ obtained from Stage 1 as a style guidance, we aim to inpaint the masked region (mostly garment) of I_m according to an input texture reference C_t^{img}. Specifically, we follow [26] and implement the inpainting by injecting the CLIP features extracted from C_t^{img} into the U-Net through cross-attention. However, a naive application of this approach does not work as the final-block CLIP features are rich in semantics but lack low-level texture, position and scale information. Addressing this issue, we propose the use of hierarchical and balanced CLIP features (Sec. 4.2) equipped with position encoding (Sec. 4.3) to learn photorealistic and non-stationary texture features, respectively:

$$C_t = \mathcal{P}_e(\text{MLP}(\mathcal{H}(C_t^{img}))) \tag{5}$$

where $\mathcal{H}$ denotes the extraction of hierarchical CLIP features, and $\mathcal{P}_e$ denotes the position encoding. We finetune a Stable Diffusion [26] model for this stage as well and have:

$$\epsilon_t = \epsilon_\theta([\mathcal{E}(P(I^m, C_s)), \mathcal{E}(I^m), z_t], t, C_t) \tag{6}$$

4.2. Photorealistic Texture Transfer

A core challenge in photorealistic texture transfer is determining optimal image features to input into the U-Net. A naive solution is to replace the text CLIP features used in previous methods with image CLIP features as images contain richer texture information. However, this does not bring significant improvement as we observed that *the low-level texture cues are gradually washed away when passing through CLIP blocks* (Fig. 3), and the resulting image CLIP features also encapsulate only high-level semantics.

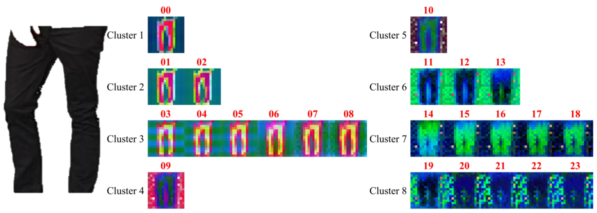

Figure 3. Visualization of clustered features from all 24 blocks of CLIP's Vision Transformer. Red: block id. The higher the block in the hierarchy, the more semantic and less texture information it contains. Please zoom in to see more details.

To solve this problem, we propose to counter the wash-away effect and achieve photorealistic texture transfer by using hierarchical and balanced CLIP features.

Hierarchical Texture Features. To incorporate low-level texture features, we propose using features extracted from all 24 blocks of CLIP's Vision Transformer. To justify this approach, we analyze these features by applying PCA (Principal Component Analysis) to generate 24 feature maps of size $16 \times 16 \times 3$. The features are then classified into eight categories through clustering. As shown in Fig. 3, there are clear differences among the eight feature clusters, indicating that hierarchical texture features provide more thorough information compared to using only the last block.

Balancing Texture Features. From Fig. 3, we observe that the features within each category are similar, while there are significant differences between categories. In addition, we find that for texture transfer, both low-level features that provide details and high-level features that provide semantic guidance are required. To minimize redundancy and balance the contributions of different types of feature, we select the shallowest layer feature from each class as the representative feature. Using this method, we obtain a representative set of hierarchical CLIP features with dimensions $(257{\times}8){\times}1024$. These features are then transformed into $(257{\times}8){\times}768$ representations via an MLP layer before being fed into the U-Net.

In this way, we obtain a minimal and balanced set of features C_t that span all levels of CLIP representations, thereby producing photorealistic results.

4.3. Learning Non-stationary Texture Features

While CLIP features excel at generating stationary textures like solid colors, they struggle with non-stationary patterns like plaids, inaccurately capturing their scales and color distributions. To address this, we add a positional encoding layer $\mathcal{P}_e$ before feeding the CLIP features C_t obtained above into the U-Net. The added positional information allows for capturing intricate visual details like plaid scales and color variations, markedly improving output quality. Thus, this enhancement substantially improves the model's ability to represent non-stationary garment textures.

Figure 4. Comparison of our model on clothing style editing. We use a garment image reference for better visualization.

5. Experiment

5.1. Experimental Setup

Datasets. While we evaluate our method on the DeepFashion Multimodal [20], SHHQ [8] and VITON-HD [6] datasets, we conduct the main experiments on the DeepFashion Multimodal dataset to enable fair comparison, as it serves as the primary benchmark across most state-of-the-art approaches.

Implementation Details. *[Stage 1]* For the DeepFashion-Multimodal dataset, we use both text and garment images as style references. For the SHHQ and VITON-HD datasets, which lack text annotations, we exclusively utilize images as inputs. The training and testing resolutions are 512×256 and $1,024 \times 512$ for DeepFashion-Multimodal and SHHQ, while we use 512×384 / $1,024 \times 768$ for VITON-HD. Each model is fine-tuned from a pre-trained stable diffusion model [26] for 50 epochs with an initial learning rate of $1e^{-5}$. The classifier-free guidance scale is set to 8 for testing. *[Stage 2]* We extract clothing from human images and change the background pixel values to 255. The training and testing resolutions match Stage 1. We fine-tune each model for 100 epochs using the same learning rate as Stage 1. The guidance scale is set to 20 for testing.

5.2. Comparison with SOTA Methods

Style Prediction Accuracy. To facilitate a fair comparison, we leverage text descriptions C_s^{text} as style guidance and evaluate all methods on 6 common clothing styles: "dress", "jumpsuit", "short sleeve top and long pants", "long sleeve top and shorts", "sleeveless top and a skirt". As shown qualitatively in Fig. 4 and quantitatively in Table 2, our method achieves significantly higher style prediction accuracy com-

Methods	Accuracy ↑	M ↑	R ↑
FashionTex [19]	82.75%	3.47%	8.33%
Text2Human [15]	88.87%	7.64%	9.73%
Ours	**92.35%**	**88.89%**	**81.94%**

Table 2. Comparison for clothing style editing. 'Accuracy': The quantitative comparison of whether the model succeeds in getting the target cloth type. 'M' and 'R' : Two user studies to objectively evaluate our methods compared to others on style fidelity and image naturalness. 'M' stands for "Match", *i.e.*, how well the result of a method matches the input style or texture; 'R' stands for "Realistic", *i.e.*, how realistic the result of a method looks.

	Texture patch		Garment	
Methods	FID ↓	KID ↓	FID ↓	KID ↓
Texture Reformer[31]	28.43	5.74	26.99	16.02
Paint-by-Example [37]	28.73	6.33	27.47	5.89
PIDM[4]	–	–	23.90	4.40
FashionTex [19]	30.11	10.99	–	–
Ours	**21.25**	**0.22**	**22.41**	**0.81**

Table 3. Quantitative Comparisons for garment texture transfer task. The KID is scaled by 1000 following [16].

pared to prior arts. In particular, our approach proficiently generates aligned images from text descriptions, whereas Text2Human [15] and FashionTex [19] struggle to produce the desired outcomes from the same textual inputs.

Texture Quality. In Table 3 and Fig. 5, we provide quantitative and visual comparisons under two different types of texture input C_t^{img}: garment texture patches and full garment images. Quantitatively, our method outperforms all SOTA methods by a large margin. Qualitatively, when using garment texture patches as C_t^{img}, Texture Reformer [31] and FashionTex [19] results exhibit high texture similarity to C_t^{img}, yet lack realism as garments. When using full garment images as C_t^{img}, PIDM [4] generates more garment-like results, but fails to retain the clothing style of the input. As for Paint-By-Example, it yields bad results on both patch and image, possibly because it inherently lacks a focus on clothing editing. However, our approach not only preserves texture fidelity, but also produces realistic, natural-looking human images in both cases.

User Study. We ran six user studies with 96 participants across various identities and age groups to objectively evaluate our methods compared to others on style/texture fidelity and image quality. Similar to [42], the percentages of each method being selected as the best one are shown in Table 2 and Table 4, which shows that our method is chosen as the best method by the majority of participants (larger than 70 %) in all cases. More details about the user study are included in the supplementary material.

	Texture patch		Garment	
Methods	M ↑	R ↑	M ↑	R ↑
Texture Reformer[31]	5.73%	1.56%	5.36%	2.98%
Paint-by-Example [37]	5.21%	6.76%	0%	4.17%
PIDM[4]	–	–	23.21%	16.66%
FashionTex [19]	1.56%	3.13%	–	–
Ours	**87.5%**	**88.55%**	**71.43%**	**76.19%**

Table 4. User studies to objectively evaluate our methods compared to others at texture fidelity and image naturalness.

5.3. Ablation Study

5.3.1 Effectiveness of Two-stage Disentanglement

To demonstrate the effectiveness of our two-stage style and texture disentanglement approach, we compare it to a one-stage variant. Specifically, this one-stage variant concatenates the CLIP features for text style input C_s^{text} and image texture input C_t^{img}, and integrates them into the U-Net of Stable Diffusion using cross-attention. Importantly, our two-stage approach enables image style input C_s^{img}, which is not possible with the one-stage variant due to the lack of paired training data for different C_s^{img}, C_t^{img}. This limitation significantly restricts the one-stage method's applicability in real-world scenarios. For a fair comparison in the limited application scenario of text style input, we remove the hierarchical and balanced feature extraction, and position encoding from our method and use only the last-block CLIP feature of C_t^{img}. As shown in Fig. 6, the one-stage model can i) accurately capture textures but be influenced by the style of the reference texture garment when extracting style information; and ii) sometimes capture the correct style but suffer from texture noise mentioned in the text (*e.g.*, plaid patterns). In contrast, our two-stage model can perfectly decouple style and texture and produce highly controllable and realistic results.

5.3.2 Effectiveness of Improved CLIP Features

To demonstrate the effectiveness of our hierarchical and balanced CLIP feature extraction, in Table 5, we present a quantitative comparison of our algorithm using CLIP features from different categories. ' Specifically, 'Single CLIP feature" refers to using only the image CLIP feature, while "layers $1 - n$" indicates concatenating the shallowest feature of the first to n^{th} category, forming a composite feature (Sec. 4.2). Experimental results show that using multi-layer features significantly outperforms single-layer features, with using shallowest feature of all 8 categories yielding the best results. As Fig. 6 shows, compared to our results, garments generated using only the Single CLIP feature do not closely match the provided texture. In contrast, our composite CLIP feature enables garment synthesis that better preserves both global style and fine texture details.

Figure 5. Visual Comparison on garment texture transfer.

Figure 6. The ablation studies of our two-stage disentanglement and hierarchical CLIP features. C_s^{text1}: *"He wears a short-sleeve T-shirt with floral patterns."* C_s^{text2}: *"His shirt has long sleeves, cotton fabric and plaid patterns."*

5.3.3 Effectiveness of Position Encoding

In Fig. 7, we conduct three experiments to analyze the effect of position encoding, including: i) without position encoding; ii) input different cropped areas from a single reference garment C_t^{img} with non-stationary textures; and iii) input C_t^{img} with the same pattern but at different scales. The results show position encoding significantly enhances the correlation between the generated garments and provided textures. Specifically, it not only i) preserves the texture distribution (*e.g.*, non-stationary) of C_t^{img} effectively, but also ii) enables generating results at different scales, introducing greater flexibility. Overall, position encoding is crucial for establishing spatial correspondence between input texture and output garments, enabling virtual try-on and fashion design that adapt to various scenarios.

Method	FID ↓	KID ↓
Single CLIP feature	24.57	4.57
layers 1 – 2	22.87	0.70
layers 1 – 4	22.99	1.04
layers 1 – 6	23.06	1.34
Ours (1 – 8)	**22.25**	**0.22**

Table 5. Effectiveness of hierarchical and balanced CLIP features.

5.4. Applications

In-shop Virtual Try-on. We qualitatively compare our virtual try-on results to state-of-the-art methods using in-shop clothing images. As Fig. 8 shows, our approach achieves comparable visual realism and clothing accuracy to recent algorithms, as evidenced by the realistic rendering and natural fit of the virtually dressed person. This demonstrates that our method can synthesize high-fidelity images of in-shop clothing on people using only a single reference image.

Fashion Design. Beyond virtual try-on, our method can also assist fashion design. As shown in Fig. 9, given a text prompt specifying the clothing style, a base texture, and a logo (we manually scale and place the logo into a proper position of I^m), our approach can generate realistic images of a person wearing the described garment. This demonstrates the versatility of our algorithm to not only virtually dress people in existing clothing, but also create new outfit designs from scratch. Our system has the potential to serve as an inspirational tool for fashion designers by instantly visualizing creative concepts.

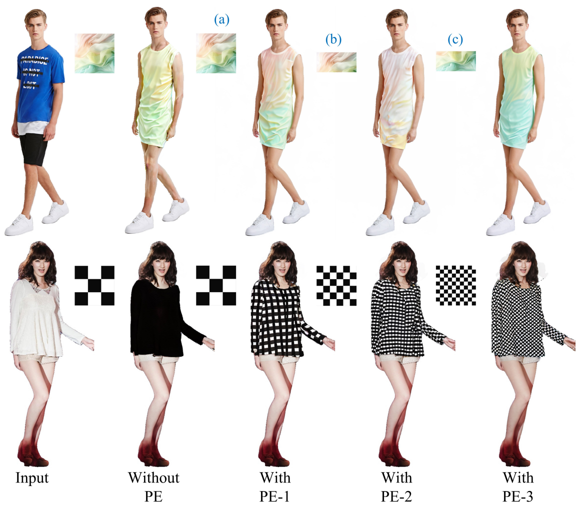

Figure 7. The effectiveness of position encoding, where (b) and (c) are the upper part and lower part of (a) respectively.

Figure 8. The comparison of virtual try-on based on in-shop cloth.

5.5. Limitations

We carefully examined our approach and identified the following limitations: i) The performance of our method relies on the accuracy of the human parsing map $P(I)$, which can be less accurate for extreme cases such as low lighting and severe occlusion. ii) Although being more general and flexible, the disentanglement and recombination of fashion elements enabled by our method inevitably contradict the "warping" paradigm of traditional VTON methods. Therefore, when VTONing full garments with structured patterns, our method introduces additional logo and texture extraction steps that are slightly more complex than existing in-shop VTON methods.

6. Conclusion

In conclusion, we have introduced the novel task of virtual try-on from unconstrained fashion designs (ucVTON) to enable flexible and photorealistic synthesis of personalized composite clothing. Our key technical contributions include a two-stage disentanglement pipeline to explicitly separate style and texture when using full garment images as complex entangled conditions;a novel hierarchical and balanced CLIP feature extraction module and position encoding to represent non-stationary textures for high-fidelity synthesis, which significantly expands the diversity of allowable style and texture conditions compared to prior arts. Extensive experiments demonstrate the superiority of our approach in photorealism, personalization, and fine-grained controllability. The flexible mixing and matching of styles and textures enabled by our work brings VTON to a new level that benefits various real-world applications from online shopping to fashion design.

Figure 9. Fashion design results under different conditions including style, texture, and design elements (*e.g*., logo).

Acknowledgements. The work was supported in part by the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen-HK S&T Cooperation Zone, Guangdong Provincial Outstanding Youth Project No. 2023B1515020055, the National Key R&D Program of China with grant No. 2018YFB1800800, by Shenzhen Outstanding Talents Training Fund 202002, by Guangdong Research Projects No. 2017ZT07X152 and No. 2019CX01X104,by Key Area R&D Program of Guangdong Province (Grant No. 2018B030338001), by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001), and by Shenzhen Key Laboratory of Big Data and Artificial Intelligence (Grant No. ZDSYS201707251409055). It is also partly supported by NSFC-62172348, 61931024 and Shenzhen General Project No. JCYJ20220530143604010.

References

[1] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *Proceedings of the IEEE/CVF conference on computer Vision and pattern recognition*, pages 18511–18521, 2022. 2

[2] Shuai Bai, Huiling Zhou, Zhikang Li, Chang Zhou, and Hongxia Yang. Single stage virtual try-on via deformable attention flows. In *European Conference on Computer Vision*, pages 409–425. Springer, 2022. 1

[3] Alberto Baldrati, Davide Morelli, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Multimodal garment designer: Human-centric latent diffusion models for fashion image editing. *arXiv preprint arXiv:2304.02051*, 2023. 2

[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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5968–5976, 2023. 3, 6

[5] Lele Chen, Justin Tian, Guo Li, Cheng-Haw Wu, Erh-Kan King, Kuan-Ting Chen, Shao-Hang Hsieh, and Chenliang Xu. Tailorgan: making user-defined fashion designs. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3241–3250, 2020. 2

[6] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14131–14140, 2021. 1, 5

[7] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, and Jian Yin. Towards multi-pose guided virtual try-on network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9026–9035, 2019. 1

[8] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen-Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. *arXiv preprint*, arXiv:2204.11823, 2022. 3, 5

[9] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8485–8493, 2021. 1

[10] Vidit Goel, Elia Peruzzo, Yifan Jiang, Dejia Xu, Nicu Sebe, Trevor Darrell, Zhangyang Wang, and Humphrey Shi. Pair-diffusion: Object-level image editing with structure-and-appearance paired diffusion models. *arXiv preprint arXiv:2303.17546*, 2023. 2

[11] Junhong Gou, Siyu Sun, Jianfu Zhang, Jianlou Si, Chen Qian, and Liqing Zhang. Taming the power of diffusion models for high-quality virtual try-on with appearance flow. *arXiv preprint arXiv:2308.06101*, 2023. 1, 2, 3

[12] Sen He, Yi-Zhe Song, and Tao Xiang. Style-based global appearance flow for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3470–3479, 2022. 1

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[14] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*, 2023. 2

[15] Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics (TOG)*, 41(4):1–11, 2022. 2, 3, 6

[16] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020. 6

[17] Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7701–7711, 2023. 2

[18] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *European Conference on Computer Vision*, pages 204–219. Springer, 2022. 1

[19] Anran Lin, Nanxuan Zhao, Shuliang Ning, Yuda Qiu, Baoyuan Wang, and Xiaoguang Han. Fashiontex: Controllable virtual try-on with text and texture. *arXiv preprint arXiv:2305.04451*, 2023. 1, 2, 3, 6

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

[21] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: High-resolution multi-category virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2231–2235, 2022. 1

[22] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. LaDI-VTON: Latent Diffusion Textual-Inversion Enhanced Virtual Try-On. In *Proceedings of the ACM International Conference on Multimedia*, 2023. 1, 2

[23] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2

[24] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 2

[25] Mengwei Ren, Mauricio Delbracio, Hossein Talebi, Guido Gerig, and Peyman Milanfar. Multiscale structure guided diffusion for image deblurring. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision, pages 10721–10733, 2023. 2

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3, 4, 5

[27] Nirat Saini, Hanyu Wang, Archana Swaminathan, Vinoj Jayasundara, Bo He, Kamal Gupta, and Abhinav Shrivastava. Chop & learn: Recognizing and generating object-state compositions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20247–20258, 2023. 2

[28] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[29] Zhengwentai Sun, Yanghong Zhou, Honghong He, and PY Mok. Sgdiff: A style guided diffusion model for fashion synthesis. *arXiv preprint arXiv:2308.07605*, 2023. 2

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[31] Zhizhong Wang, Lei Zhao, Haibo Chen, Ailin Li, Zhiwen Zuo, Wei Xing, and Dongming Lu. Texture reformer: towards fast and universal interactive texture transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2624–2632, 2022. 2, 3, 6

[32] Qiucheng Wu, Yujian Liu, Handong Zhao, Trung Bui, Zhe Lin, Yang Zhang, and Shiyu Chang. Harnessing the spatial-temporal attention of diffusion models for high-fidelity text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7766–7776, 2023. 2

[33] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 3

[34] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2256–2265, 2021. 2

[35] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black. Econ: Explicit clothed humans optimized via normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 512–523, 2023. 2

[36] Han Xue, Zhiwu Huang, Qianru Sun, Li Song, and Wenjun Zhang. Freestyle layout-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14256–14266, 2023. 2

[37] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023. 2, 6

[38] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7850–7859, 2020. 1

[39] Ruiyun Yu, Xiaoqi Wang, and Xiaohui Xie. Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10511–10520, 2019. 1

[40] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2

[41] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *arXiv preprint arXiv:2305.16322*, 2023. 2

[42] Luyang Zhu, Dawei Yang, Tyler Zhu, Fitsum Reda, William Chan, Chitwan Saharia, Mohammad Norouzi, and Ira Kemelmacher-Shlizerman. Tryondiffusion: A tale of two unets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4606–4615, 2023. 6