M&M VTO: Multi-Garment Virtual Try-On and Editing

Luyang Zhu[1,2*]   Yingwei Li[1]   Nan Liu[1]   Hao Peng[1]

Dawei Yang[1]   Ira Kemelmacher-Shlizerman[1,2]

[1]Google Research   [2]University of Washington

Figure 1. Given an input person image, multiple garments, M&M VTO can output a virtual try-on visualization of how those garments would look on the person. Our model performs well across various body shapes, poses, and garments. In addition, it allows layout to be changed, *e.g.*, "roll up the sleeves" (top rightmost column), and "tuck in the shirt and roll down the sleeves" (bottom rightmost column).

Abstract

We present M&M VTO–a mix and match virtual try-on method that takes as input multiple garment images, text description for garment layout and an image of a person. An example input includes: an image of a shirt, an image of a pair of pants, "rolled sleeves, shirt tucked in", and an image of a person. The output is a visualization of how those garments (in the desired layout) would look like on the given person. Key contributions of our method are: 1) a single stage diffusion based model, with no super resolution cascading, that allows to mix and match multiple garments at $1024{\times}512$ resolution preserving and warping intricate garment details, 2) architecture design (VTO UNet Diffusion Transformer) to disentangle denoising from person specific features, allowing for a highly effective finetuning strategy for identity preservation (6MB model per individual vs 4GB achieved with, e.g., dreambooth finetuning); solving a common identity loss problem in current virtual try-on methods, 3) layout control for multiple garments via text inputs

[1]Work done while author was an intern at Google.

finetuned over PaLI-3 [8] for virtual try-on task. Experimental results indicate that M&M VTO achieves state-of-the-art performance both qualitatively and quantitatively, as well as opens up new opportunities for virtual try-on via language-guided and multi-garment try-on.

1. Introduction

Virtual try-on (VTO) is the task of synthesizing how a person would look in various garments based on provided garment photos and a person photo. Ideally the synthesis is high resolution, showcasing the intricate details of garments, while at the same time representing the body shape, pose, and identity of the person accurately. In this paper, we focus specifically on multiple garment VTO and editing. For example, a user of our method would provide one or more photos for garments, e.g., shirt, pants, and one photo of a person, with additional optional text input to request a layout, e.g., "shirt tucked out, rolled sleeves". The garment photos could be either a product photo (layflat) or the garment as worn by a different person. The person photo would be a full field-of-view photo showing the person head to toe. Our method, which we named, M&M VTO, outputs a visualization of how the person looks in those garments. Figure 1 shows a couple of examples.

Redefining the VTO problem as multiple-garment VTO, rather than the commonly targeted single garment VTO, allowed us to deeply rethink architecture design and solve several open problems in multi, as well as single VTO networks, in addition to opening up the new possibilities for mix and match and editing layouts.

Two of the most challenging VTO problems are (1) how to preserve the small but important details of garments while warping the garment to match various body shapes, and (2) how to preserve the identity of the person without leaking the original garments that the person was wearing to the final result. State-of-the-art methods came close for single garment VTO by leveraging the power of diffusion, and building networks that denoise while warping, e.g., Parallel-Unet [59]. To address (1), however, the network requires to max out the number of parameters and a memory heavy Parallel-UNet to warp a single garment. For (2) a "clothing-agnostic" representation is typically used for the person image to erase the current garment to be replaced by VTO, but at same time it removes a significant amount of identity information, with the network needing to hallucinate the rest, resulting in loss of characteristics like tattoos, body shape or muscle information.

With more garments, as in multi-garment VTO, the number of pixels needed to go through the network triples, so the same number of parameters would create a lower quality VTO. Similarly, showing head to toe person and allowing multiple garments, means 'clothing-agnostic' represen-

tation leaves even less of the identity of the person–if just a shirt needs to be replaced, the network can still see how the bottom part of that person looks like (and shape of the legs), while if all garments are changing the agnostic would preserve even less information about the person.

Our solution, M&M VTO, is three-fold as depicted in Figure 2. First, we designed a single-stage diffusion model to directly synthesize 1024×512 images with no need for extra super-resolution(SR) stages as commonly done by state-of-the-art image generation techniques. We found that as we expand the scope of VTO, having cascaded design is detrimental as the base model's low resolution assumes excessive downsampling of ground truth during training, thus losing forever garment details; as SR models depend heavily on the base model, if the details disappear they can not be upsampled effectively. Training a single stage base model just on higher resolution data, however, does not solve the problem, as the model doesn't converge even with ideas proposed in [7, 23]. Instead we designed a progressive training strategy where model training begins with lower-resolution images and gradually moves to higher-resolution ones during the single stage training. Such a design naturally benefits training at higher resolutions by utilizing the prior learned at lower resolutions, allowing the model to better learn and refine high-frequency details.

Second, to solve the identity loss (and/or clothing leakage) during the 'clothing-agnostic' process, we propose a space saving finetuning strategy. Rather than finetuning the entire model during post processing, as commonly done by techniques like DreamBooth [44], we choose to finetune person features only. We designed a VTO UNet Diffusion Transformer (VTO-UDiT) to isolate encoding of person features from the denoising process. In addition to producing much higher quality results, this design also drastically reduces finetuned model size per new individual, going from 4GB to 6MB.

Third, we created text based labels representing various garment layout attributes, e.g., rolled sleeves, tucked in shirt, and open jacket. We formulated attribute extraction as an image captioning task and finetuned a PaLI-3 model [8] using only $1.5k$ labeled images. This allows us to automatically extract accurate labels for the whole training set.

Above three design choices are critical in producing high quality VTO results for multi-garment scenarios. We perform detailed ablation studies, and comparisons to state-of-the-art papers to illustrate each design choice. Our method significantly outperforms others. The user study shows that our method is chosen as best 78.5% of the time compared to state-of-the-art on multiple-garment VTO task.

2. Related Work

In this section we will focus on related work relevant to our three key design choices described above. For a compre-
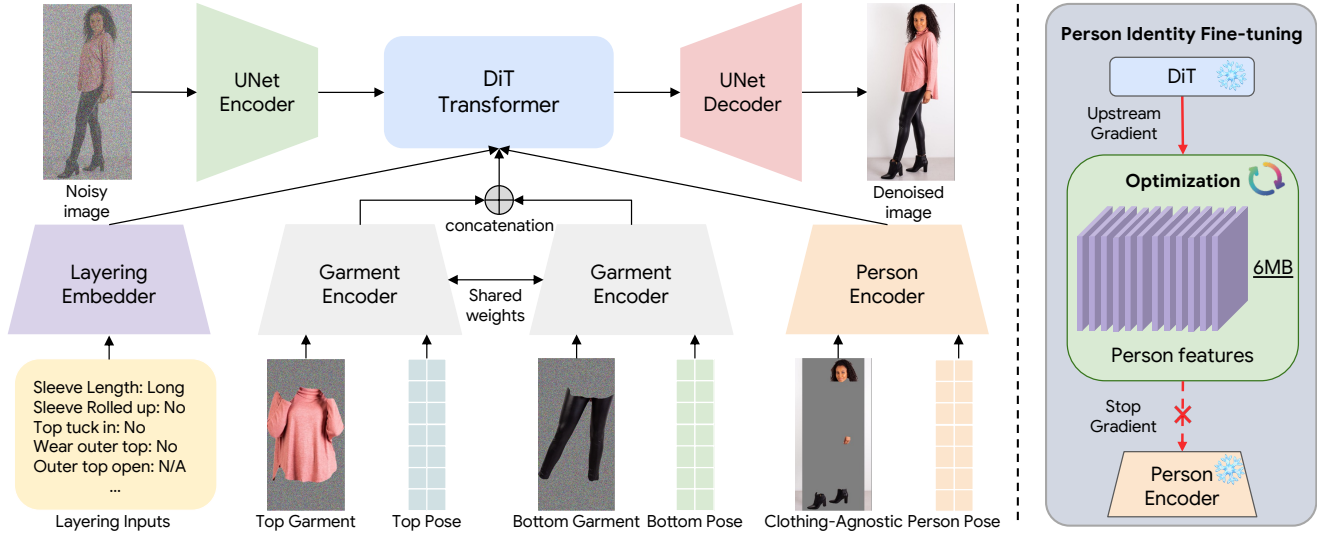
Figure 2. **Overview of M&M VTO. Left**: Given multiple garments (top and bottom in this case, full-body garment not shown for this example), layout description, and a person image, our method enables multi-garment virtual try-on. **Right**: By freezing all the parameters, we optimize person feature embeddings extracted from the person encoder to improve person identity for a specific input image. The fine-tuning process recovers the information lost via agnostic computation.

hensive list of recent papers in virtual try-on we also invite the reader to review this list[1].

Image-Based Virtual Try-On. The seminal VITON method [17] proposed a warping model that estimates pixel displacements between the original garment image and target warp. Based on those displacements, it warped the garment, and then used a blending model to combine the warped garment with the person image, showing one of the first promising results for VTO. Many works followed, to improve pixel displacement estimation. [52] proposed thin plate splines, [58] predicted target segmentation and parsing for improved warping, student-teacher approach and distillation were proposed by [16, 24]. Other efforts include adaptive parsing and second order constraint on thin plate splines [56], optimization to remove misalignments [9], leveraging dance videos to improve warping [12], regularizing [57], and using self and cross attention to improve flow computation [3]. With the rise of Style-GAN, [18] proposed StyleGAN for optical flow, [29] proposed a generator-discriminator approach, [32, 54, 55] reported improved results for flow compute and inpainting by utilization of landmarks, and [6] incorporated size information.

While results were improving, there was an inherent difficulty in warping garments *explicitly–pixel wise*, as there is too much variation in folds, logos, texture where a garment image needs to warp to a new body shape. Rather than estimating flows directly, [30] proposed to interpolate

StyleGAN coefficients to create try-on, still lacking complex textures, though, due to the averaging nature of Style-GAN. TryOnDiffusion [59] introduced a diffusion-based [21, 48, 50] Parallel-UNet enabling implicit warping and blending in the same model via cross-attention, showing significantly better results. Key limitations of that approach were incomplete garment details due to base model being only $128{\times}128$ resolution, and identity preservation. Finally, most of those methods are focused on single garment try-on only.

Finetuning Diffusion. As finetuning is a general concept and wasn't much used for VTO, we will review recent works for any general finetuning. Sometimes also called personalization [14], finetuning is the task of adjusting an existing, say text to image generation model, to a specific task, e.g., style transfer. Dreambooth [44] showed fantastic results by finetuning on a few images, and accompanying text, to bind a unique identifier with a specific subject. [15, 31] learned encoders to transfer visual concept into textual embeddings. [1] created a network that maps noise timestamp and layer to text token space. To improve multi-concept composition [33], Custom Diffusion [28] optimized concept embeddings along with key and value projection matrices of cross attention layers in the text-to-image model. In contrast, our approach is tailored for VTO and requires only **6**MB of parameters per person during the inference phase.

Image Editing with Diffusion Models. Editing of general images with diffusion initially utilized image masks [2, 10,
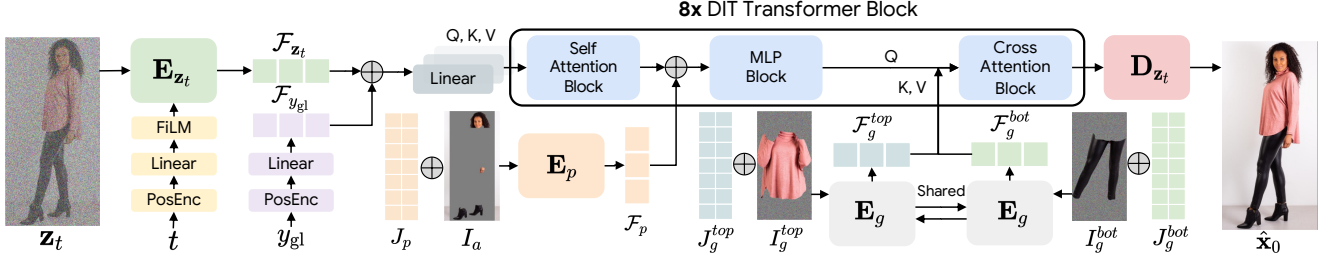

Figure 3. VTO-UDiT architecture. For image inputs, UNet encoders ($\mathbf{E}_{\mathbf{z}_t}$, $\mathbf{E}_p$, $\mathbf{E}_g$) extract features maps ($\mathcal{F}_{\mathbf{z}_t}$, $\mathcal{F}_p$, $\mathcal{F}_g^\kappa$) from $\mathbf{z}_t$, I_a, I_c^κ, respectively, with $\kappa \in \{\text{upper}, \text{lower}, \text{full}\}$. Diffusion timestep t and garment attributes y_{gl} are embedded with sinusoidal positional encoding, followed by a linear layer. The embeddings ($\mathcal{F}_t$ and $\mathcal{F}_{y_{\text{gl}}}$) are then used to modulate features with FiLM [13] or concatenated to the key-value feature of self-attention in DiT similar to [46]. Following [59], spatially aligned features($\mathcal{F}_{\mathbf{z}_t}$, $\mathcal{F}_p$) are concatenated whereas $\mathcal{F}_g^\kappa$ are implicitly warped with cross-attention blocks. The final denoised image $\hat{\mathbf{x}}_0$ is obtained with decoder $\mathbf{D}_{\mathbf{z}_t}$, which is architecturally symmetrical to $\mathbf{E}_{\mathbf{z}_t}$.

34, 35, 38, 45]. SDEdit [35] added noise to the inputs and then subsequently denoised them through a stochastic process. Palette [45] trained a conditional diffusion model for specific edit tasks. BlendedDiffusion [2], inspired by CLIP guided diffusion [11], utilized CLIP text encoder [41] and spatial masks to edit images by blending noised input images with locally generated contents. Requiring masks is not applicable to VTO tasks e.g., tuck this shirt in.

The success of text to image diffusion models [22, 38, 42, 43] led to text-based image editing [5, 10, 19, 26, 27, 36, 47, 51, 53]. For example, DiffEdit [10] infers a region mask based on text instructions, and then guides image editing using inverted noise resulted from DDIM inversion process [49]. Prompt-to-Prompt (P2P) [19] edits images using only text by manipulating the cross-attention scores conditioned on inverted latents. Null-text inversion [36] optimized on null-text embeddings by minimizing differences between latent codes from unconditional inversion process and conditional one. InstructPix2Pix [5] directly manipulates image in the denoising process by using finetuned Stable Diffusion trained on paired examples generated using P2P technique with given editing instructions. Generally, text based editing, while allowing for easier input (compared to masks), often creates the edit but fails to preserve original image details, e.g., in VTO case the original garment details are lost with such techniques. We solve it via VTO specific finetuning on PaLI-3 and then using it as condition in the network.

3. Method

Given a person image I_p, an upper-body garment image I_g^{upper}, a lower-body garment image I_g^{lower} and a full-body garment image I_g^{full}, our method synthesizes VTO result I_{tr} for person p. Optionally, a layout attribute is provided as input as well. We begin by describing training data and its preprocessing, and then the model design of M&M VTO.

3.1. Dataset Preparation and Preprocessing

M&M VTO is trained on pairs–person image I_p, and a garment image I_g. I_g can be an image of a garment laid out on a flat surface (layflat), or an image of a person wearing the garment (most often in another pose). As the pair assumes that they share only one or two garments rather than all three of upper, lower and full, we do the following simple process. We compute a garment embedding for each of the three garments (determined by segmentation) and compare which one appears on the person image. The ones that do not are set to 0.

Each pair is then processed following [59]. Conditional inputs $\mathbf{c}_{\text{tryon}}$ includes clothing-agnostic RGB I_a, segmented garment I_c^κ, 2D pose keypoints J_p for the person image I_p and 2D pose keypoints J_g^κ for garment images I_g^κ (J_g^κ is a vector with all -1's if I_g^κ is a layflat garment image). To make sure that background is as tight as possible (allowing for the model to fully focus on garments) we crop and resize all images to 1024×512, approximately resembling aspect ratio of a photograph of a head to toe person.

We also introduce a layout input y_{gl}, defining desired attributes of the garments. We only focus on attributes that one can do in real-life, for example: roll up sleeves, tuck in the shirt, etc. rather than changing texture or garment properties. Full set of attributes is in the supplementary material. One way to calculate attributes of each garment is by training a classifier for each attribute. We chose instead to finetune a large vision language model (PaLI-3[8]). Specifically, we convert all attributes into a formatted text and formulate it as an image captioning task. There are two advantages for this formulation. First, vision language models have strong priors trained on large datasets and can utilize the correlation between different garment layout attributes (e.g. the sleeve can not be rolled up if the sleeve type is sleeveless). Second, using a single model can also accelerate the training data generation process. Thanks to

Figure 4. **Qualitative Comparison with existing Try-On methods.** On the left, we compare with TryOnDiffusion [59] on our test set and further evaluate on DressCode [37] dataset, as shown on the right. Our method can generate better garment details and layouts.

the strong prior encoded in the PaLI-3 model, we are able to get very accurate garment attributes by finetuning PaLI-3 with only 1,500 images. To get y_{gl} for each training sample, we first extract garment layout attributes relevant to the garment type κ by running finetuned PaLI-3 on I_g^κ, and then concatenate those attributes into a single vector. Refer to the supplementary for more details.

3.2. Single Stage M&M VTO

Cascaded diffusion models, i.e., lower resolution diffusion base model, followed by super resolution models, have shown great success for text to synthetic image generation [22, 53]. Similarly, for VTO [59] followed a similar setup where three stages were used. For multi-garment VTO, however, such design is performing poorly, as the base model doesn't have enough capacity to create intricate warps and occlusions based on person's body shape. We observed that high-frequency garment details are smoothed and blurred out if images are downsampled by more than 2 times. Thus, it is impossible for base diffusion models trained to preserve those garment details as their groundtruth images do not include them.

Ideally we would just synthesize 1024×512 images with the base model directly. This turned out to be a challenging task, as if the cross-attention is applied at a lower resolution, the high frequency image details are destroyed by excessive downsampling of feature maps, and the model tends to learn a global structure for the warping. On the other hand, applying cross-attention at a higher resolution does not converge under random initialization from our initial experiments.

To tackle this challenge, we use an effective progressive training paradigm for M&M VTO. The key idea is to initialize the higher resolution diffusion models using a pre-trained lower resolution one. Specifically, we first train a base diffusion model to synthesize 512×256 try-on results $I_{tr}^{512\times256}$, where the cross-attention happens in 32×16. After that, we continue to train the *exact same model* to synthesize 1024×512 try-on results $I_{tr}^{1024\times512}$, where the cross-attention happens in 64×32 with the same architecture. Note that our training algorithm does not require modifying or adding new components to the architecture, all we need is to train the model with data in different resolutions, which is easy to implement.

3.3. VTO-UDiT Architecture

The VTO-UDiT network (Figure 3) is represented as

$$\hat{\mathbf{x}}_0 = \mathbf{x}_\theta(\mathbf{z}_t, t, \mathbf{c}_{tryon}) \tag{1}$$

where t is the diffusion timestep, $\mathbf{z}_t$ is the noisy image corrupted from the ground-truth $\mathbf{x}_0$ at timestep t, $\mathbf{c}_{tryon}$ is the try-on conditional inputs, and $\hat{\mathbf{x}}_0$ is the predicted clean image at timestep t. In practice, we follow [22] to set the network output in $\mathbf{v}$-space to avoid color drift issues in higher resolution diffusion models. Given the predicted $\hat{\mathbf{v}}_t$, we compute $\hat{\mathbf{x}}_0 = \alpha_t \mathbf{z}_t - \sigma_t \hat{\mathbf{v}}_t$, where $\alpha_t, \sigma_t \in (0,1)$ control the signal-to-noise ratio.

Inspired by [23], we change the Parallel-UNet architecture [59] into a UDiT architecture where the transformer block is implemented as DiT [39]. With the combination of UNet and DiT, the model benefits from light weight UNet as image encoders and the heavy DiT blocks to process in lower resolution feature maps for attention operations.
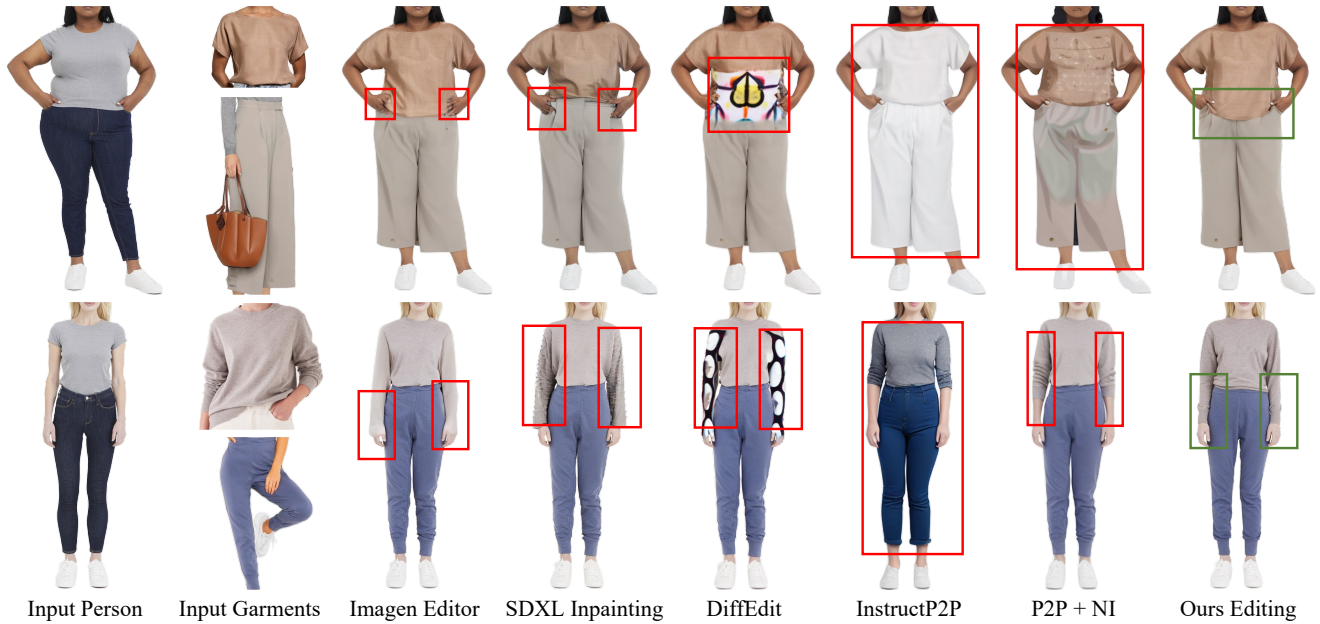
| Input Person | Input Garments | Imagen Editor | SDXL Inpainting | DiffEdit | InstructP2P | P2P + NI | Ours Editing |

Figure 5. **Qualitative Comparison for Garment Layout Editing.** Top: editing instruction is to "tuck out the shirt". Bottom: "roll down the sleeve". Our method enables more accurate layout editing while preserving the details from the inputs. Details are provided in the Supplementary.

Moreover, the design of UDiT fully disentangle the encoding process of $\mathbf{c}_{\text{tryon}}$ from the denoising process, which is critical for person feature finetuning described later in Section 3.4. More specifically, 1) Different UNet encoders are used to process the input images without information exchange. 2) Only $\mathbf{E}_{\mathbf{z}_t}$ takes diffusion timestep t embedding as as input, while $\mathbf{E}_p$ and $\mathbf{E}_g$ do not, to fully disentangle conditional features from diffusion denoising. 3)Unlike Parallel-UNet [59] which updates both conditional features and noisy image features in parallel, VTO-UDiT fixes the conditional features and only updates diffusion features during the forward pass of DiT blocks.

Also, note that all UNet encoders are fully convolutional and free of attention operations, which is preferable for progressive training mentioned in Section 3.2.

3.4. Efficient Finetuning for Person Identity

A key challenge of current VTO methods is the loss of person identity due to the use of clothing-agnostic representation. To tackle this problem, we propose a space-efficient finetuning strategy based on our VTO-UDiT architecture. As described in Sec. 3.3, person feature $\mathcal{F}_p$ is independent of diffusion or garment related features, and is kept fixed for DiT blocks where conditioning happens. Thus, we are able to directly finetune the person features instead of the whole diffusion model. This greatly reduces the optimizable weights from 4GB to 6MB. Furthermore, we found finetuning on person features will not cause the model to

overfit on the particular garments worn by the target person as shown in Section 4.

The finetuning process needs to learn how to warp garments from varying sizes and poses on the target person, however, acquiring pairs of images of same garment and various shapes and sizes is impractical. Instead we use pretrained M&M VTO to prepare a synthetic dataset. We segment out garments worn by the target person image, and try-on the garment on multiple person images across various poses (*e.g.* different torso orientations and arm positions) and body shapes (from 2XS to 2XL), resulting in 150 samples. Since our pretrained M&M VTO can accurately preserve but warp garment details to new pose and shape, the quality of the synthetic finetuning data is high, and allows us to reconstruct the person identity when tested on unseen garments.

4. Experiments

In this section, we describe datasets, comparisons and ablations. Additional results as well as implementation details can be found in supplementary.

Datasets. Our model is trained on two types of datasets: 1) "garment paired" dataset of 17 Million samples, where each sample consists of two images of the same garment in two different poses/body shapes, 2) "layflat paired" dataset of 1.8 Million samples, where each sample consists of an image with garment laid out on a flat surface and an image of a person wearing the garment. For testing, we use two

Test datasets	Ours 8, 300			DressCode		
Methods	FID ↓	KID ↓	US ↑	FID ↓	KID ↓	US ↑
GP-VTON [54]	N/A	N/A	N/A	38.392	33.909	1327
TryOnDiffusion [59]	19.459	17.617	1526	15.944	5.363	951
Ours	**18.145**	**15.227**	**6512**	**14.019**	**2.772**	**2945**
Hard to tell	N/A	N/A	262	N/A	N/A	177

Table 1. **Quantitative Comparison.** We evaluate on our 8, 300 triplets test set and DressCode triplets test set. GP-VTON [54] is trained on layflat garments, thus we report only on DressCode test set. The metrics are FID, KID, and user study (US). All baselines are run twice sequentially, first for tops then for bottoms try-on (See Section 4).

sets: 1) we collected 8, 300 triplets (top, bottom, person) that are *unseen* during training, 2) we use DressCode [37] just for comparison with other methods that use it.

Comparison of VTO. We compared with two representative state-of-the-art methods: TryOnDiffusion [59], and GP-VTON [54]. Other methods don't provide code at the time of submission. Our 8, 300 triplets test set was used to compare to TryonDiffusion, and DressCode triplets unpaired test set was used to compare to both GP-VTON and TryonDiffusion. As TryOnDiffusion was trained only on tops, and person images, we retrained it on our dataset for upper-body, lower-body, and full-body garments separately. For GP-VTON, we used officially released checkpoints trained on DressCode. Then we ran inference sequentially first to produce top VTO, and then bottom VTO. Figure 4 shows that M&M VTO outperforms baselines in aspects such as garment interactions, warping, and detail preservation. Table 1 shows that our method outperforms baselines for FID [20], KID [4] (scaled by 1000 following [25]) and user study (US). In the user study, 16 non-experts were asked to either select the best result or opt for "hard to tell." The findings indicate that users generally prefer M&M VTO over other methods. We provide results for single garment try-on, and other comparisons in supplementary.

Comparison of Editing. We evaluate our approach by comparing with several text-guided image editing methods. Inpainting mask free: Prompt-to-Prompt (P2P) [19] + Null inversion [36] (P2P + NI) and InstructPix2Pix (IP2P) [5] using a target text prompt and an input image that we wish to perform editing on. With inpainting mask: Imagen editor [53], DiffEdit [10] and SDXL inpainting [40]. Figure 5 demonstrates that our method can interpret garment layout concepts more effectively, allowing for more precise edits of the targeted part without affecting other areas. We provide quantitative comparison and additional details about specific prompts and input masks in supplementary.

Finetuning Comparison. We compare to three baselines: non-finetuned model, finetuning the full model and finetuning the person encoder. For the latter two baselines, we have incorporated the class-specific prior preservation loss,

Figure 6. **Ablation Comparison**. We provide qualitative zoom-in visualization to compare our progressive training with cascaded models and the model trained from scratch. Our approach can generate better garment details, *i.e.*, more accurate texts. See supplementary for full images.

as utilized in DreamBooth [44], to prevent overfitting to the clothing worn by the target person. For our approach, we don't apply such regularization technique as we found our method does not suffer from overfitting. Figure 7 showcases that our method successfully retains characteristics of the human models (*e.g.*, body shape) without compromising the details of the garments.

Ablation for Single Stage Model vs. Cascaded. Our method generates $1024{\times}512$ try-on images in a single stage. For the cascaded variant, we trained a $512{\times}256$ base diffusion model, followed by a $512{\times}256 \rightarrow 1024{\times}512$ SR diffusion model. Both models share the same architecture as our single-stage model, with the distinction that the SR model concatenates the low-resolution image to the noisy image. Figure 6 illustrates that our single stage model excels at maintaining complex garment details like tiny texts or logos.

Ablation for Progressive Training vs. Training from Scratch. We train an identical model from scratch on $1024{\times}512$ data, without leveraging any model pretrained in lower resolutions. Figure 6 highlights that our progressively trained model more effectively manages garment warping under significant pose variations, whereas the ablated version struggles with learning implicit garment warping through cross-attention.

Limitations. Firstly, our approach isn't designed for layout editing tasks, such as "Open the outer top." As demonstrated in Figure 8 (left), a random shirt is generated by the model, as no specific information is provided from inputs about what should be inpainted in the open area. Secondly, our method struggles with uncommon garment combinations found in the real world, like a long coat paired with skirts. As shown in the right example of Figure 8, the model tends to split the long coat in an attempt to show the skirts, because it learned from examples where both garments are typically visible during training. Thirdly, our model faces challenges when dealing with upper-body clothing from different images, *e.g.* pairing a shirt from one photo with an outer coat from another. This issue mainly stems from the difficulty in finding training pairs where one image clearly shows a shirt without any cover, while another displays the

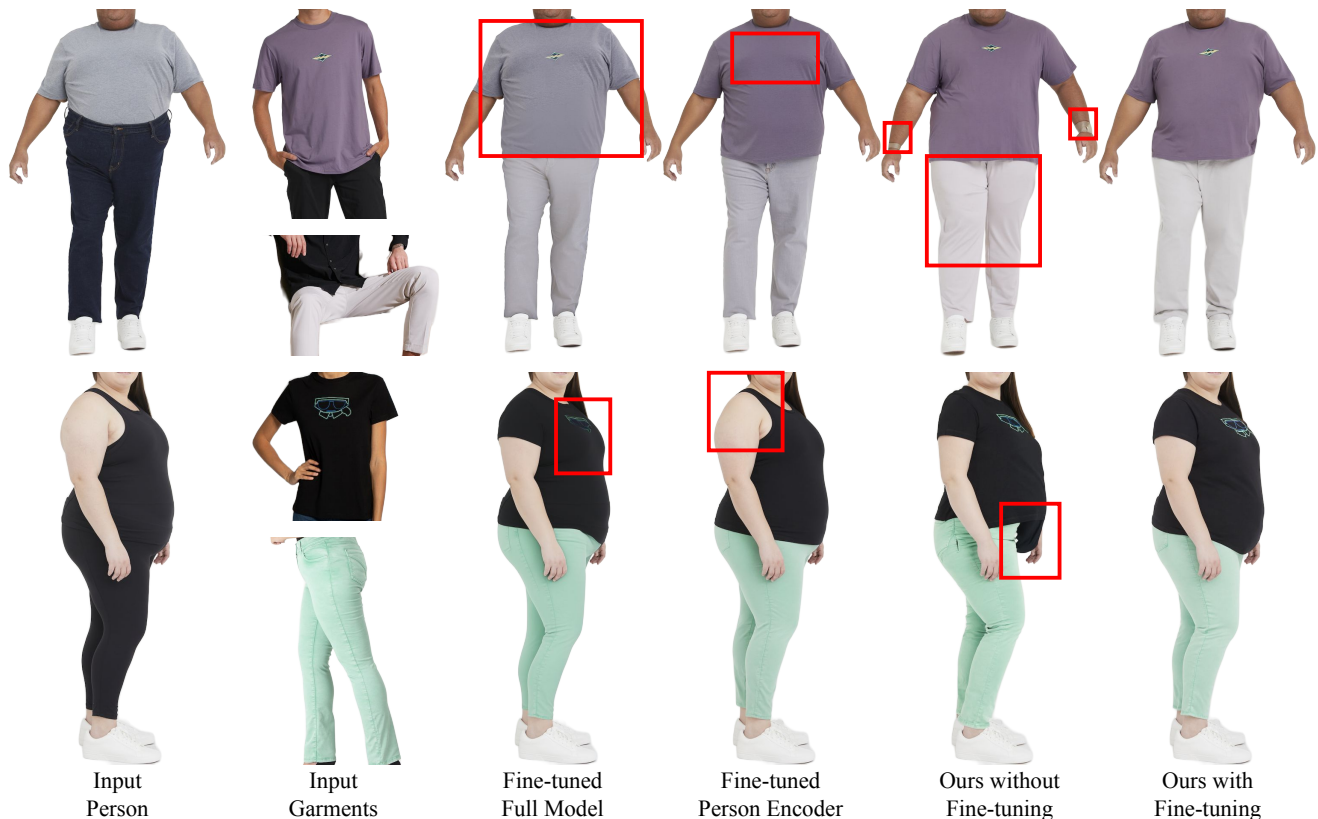| Input Person | Input Garments | Fine-tuned Full Model | Fine-tuned Person Encoder | Ours without Fine-tuning | Ours with Fine-tuning |

Figure 7. **Qualitative Comparison on Person Fine-tuning Strategy.** We provide a comparison with various types of fine-tuning strategies. Our method shows a better person identity preservation than fine-tuning the whole model or person encoder only. Red boxes highlight example errors, *e.g.*, sleeves too short, and extra fabric.

same shirt under an outer layer. As a result, the model struggles to accurately remove the shirt when it's covered by an outer layer during testing. Finally, note that our method visualizes how an item might look on a person, accounting for their body shape, but it doesn't yet include size information nor solves for exact fit.

5. Conclusion

We present a method that can synthesize multi-garment try-on results given an image of person and images of upper-body, lower-body and full-body garments. Our novel architecture VTO-UDiT as well as progressive training strategy, enabled better than state-of-the-art results, particularly in preserving fine garment details and person identity. Furthermore, our method allows for explicit control of garment layout via conditioning the model with garment attributes obtained from a finetuned vision-language model.

References

[1] Yuval Alaluf, Elad Richardson, Gal Metzer, and Daniel Cohen-Or. A neural space-time representation for text-to-image personalization, 2023. 3

Figure 8. **Failure Cases.** Our model could generate random clothing given layout information. As shown in the left example, given "outer top open", the model generates a random inner top. In addition, the model could lead to failures when dealing with rare garment combinations. For example, given a long coat and skirt combination, it creates a half open coat, shown in the right image.

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 3, 4

[3] 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. 3

[4] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 7

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 4, 7

[6] Chieh-Yun Chen, Yi-Chung Chen, Hong-Han Shuai, and Wen-Huang Cheng. Size does matter: Size-aware virtual try-on via clothing-oriented transformation try-on network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7513–7522, 2023. 3

[7] Ting Chen. On the importance of noise scheduling for diffusion models. *arXiv preprint arXiv:2301.10972*, 2023. 2

[8] Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, et al. Pali-3 vision language models: Smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*, 2023. 2, 4

[9] 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. 3

[10] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 3, 4, 7

[11] Katherine Crowson. Clip guided diffusion hq 256x256. https://colab.research.google.com/drive/12a_Wrfi2_gwwAuN3VvMTwVMz9TfqctNj. 4

[12] Xin Dong, Fuwei Zhao, Zhenyu Xie, Xijin Zhang, Daniel K Du, Min Zheng, Xiang Long, Xiaodan Liang, and Jianchao Yang. Dressing in the wild by watching dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3480–3489, 2022. 3

[13] Vincent Dumoulin, Ethan Perez, Nathan Schucher, Florian Strub, Harm de Vries, Aaron Courville, and Yoshua Bengio. Feature-wise transformations. *Distill*, 3(7):e11, 2018. 4

[14] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022. 3

[15] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. 2023. 3

[16] 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. 3

[17] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7543–7552, 2018. 3

[18] 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 (CVPR)*, pages 3470–3479, 2022. 3

[19] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 4, 7

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3

[22] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 4, 5

[23] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. *arXiv preprint arXiv:2301.11093*, 2023. 2, 5

[24] Thibaut Issenhuth, Jérémie Mary, and Clément Calauzènes. Do not mask what you do not need to mask: a parser-free virtual try-on. In *European Conference on Computer Vision*, pages 619–635. Springer, 2020. 3

[25] 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. 7

[26] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 4

[27] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022. 4

[28] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 3

[29] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *Proceedings of the European conference on computer vision (ECCV)*, 2022. 3

[30] Kathleen M Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. Tryongan: Body-aware try-on via layered interpolation. *ACM Transactions on Graphics (TOG)*, 40(4):1–10, 2021. 3

[31] Dongxu Li, Junnan Li, and Steven C.H. Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. 2023. 3

[32] Zhi Li, Pengfei Wei, Xiang Yin, Zejun Ma, and Alex C Kot. Virtual try-on with pose-garment keypoints guided inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22788–22797, 2023. 3

[33] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022. 3

[34] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 4

[35] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 4

[36] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 4, 7

[37] 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. 5, 7

[38] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 4

[39] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 5

[40] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 7

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 4

[42] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 4

[43] 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. 4

[44] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022. 2, 3, 7

[45] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 4

[46] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 2022. 4

[47] Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. Knn-diffusion: Image generation via large-scale retrieval. *arXiv preprint arXiv:2204.02849*, 2022. 4

[48] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 3

[49] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 4

[50] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[51] Dani Valevski, Matan Kalman, Yossi Matias, and Yaniv Leviathan. Unitune: Text-driven image editing by fine tuning an image generation model on a single image. *arXiv preprint arXiv:2210.09477*, 2022. 4

[52] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 589–604, 2018. 3

[53] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18359–18369, 2023. 4, 5, 7

[54] Zhenyu Xie, Zaiyu Huang, Xin Dong, Fuwei Zhao, Haoye Dong, Xijin Zhang, Feida Zhu, and Xiaodan Liang. Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23550–23559, 2023. 3, 7

[55] Keyu Yan, Tingwei Gao, Hui Zhang, and Chengjun Xie. Linking garment with person via semantically associated landmarks for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17194–17204, 2023. 3

[56] 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. 3

[57] Han Yang, Xinrui Yu, and Ziwei Liu. Full-range virtual try-on with recurrent tri-level transform. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3460–3469, 2022. 3

[58] 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. 3

[59] 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 (CVPR)*, pages 4606–4615, 2023. 2, 3, 4, 5, 6, 7