Smart Fitting Room: A One-stop Framework for Matching-aware Virtual Try-on

Mingzhe Yu
Shandong University
Jinan, China
mz_y@mail.sdu.edu.cn

Yunshan Ma
National University of Singapore
Singapore
yunshan.ma@u.nus.edu

Lei Wu*
Shandong University
Jinan, China
i_lily@sdu.edu.cn

Kai Cheng
Shandong University
Jinan, China
k_c@mail.sdu.edu.cn

Xue Li
Jiangnan University
Wuxi, China
xue.lii754@gmail.com

Lei Meng
Shandong University
Jinan, China
lmeng@sdu.edu.cn

Tat-Seng Chua
National University of Singapore
Singapore
dcscts@nus.edu.sg

ABSTRACT

The development of virtual try-on has revolutionized online shopping by allowing customers to visualize themselves in various fashion items, thus extending the in-store try-on experience to the cyber space. Although virtual try-on has attracted considerable research initiatives, existing systems only focus on the quality of image generation, overlooking whether the fashion item is a good match to the given person and clothes. Recognizing this gap, we propose to design a one-stop Smart Fitting Room, with the novel formulation of matching-aware virtual try-on. Following this formulation, we design a Hybrid Matching-aware Virtual Try-On Framework (HMaVTON), which combines retrieval-based and generative methods to foster a more personalized virtual try-on experience. This framework integrates a hybrid mix-and-match module and an enhanced virtual try-on module. The former can recommend fashion items available on the platform to boost sales and generate clothes that meets the diverse tastes of consumers. The latter provides high-quality try-on effects, delivering a one-stop shopping service. To validate the effectiveness of our approach, we enlist the expertise of fashion designers for a professional evaluation, assessing the rationality and diversity of the clothes combinations and conducting an evaluation matrix analysis. Our method significantly enhances the practicality of virtual try-on. The code is available at https://github.com/Yzcreator/HMaVTON.

*Corresponding author.

ICMR '24, June 10–14, 2024, Phuket, Thailand

ACM ISBN 979-8-4007-0619-6/24/06
https://doi.org/10.1145/3652583.3658064

CCS CONCEPTS

• **Information systems → Multimedia and multimodal retrieval**; • **Computing methodologies → Computer vision**; Machine learning.

KEYWORDS

Mix-and-match, Fashion Image Generation, Virtual Try-on

ACM Reference Format:
Mingzhe Yu, Yunshan Ma, Lei Wu, Kai Cheng, Xue Li, Lei Meng, and Tat-Seng Chua. 2024. Smart Fitting Room: A One-stop Framework for Matching-aware Virtual Try-on. In *Proceedings of the 2024 International Conference on Multimedia Retrieval (ICMR '24), June 10–14, 2024, Phuket, Thailand*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3652583.3658064

1 INTRODUCTION

Virtual try-on facilitates customers in seamlessly envisioning their appearance while perusing diverse fashion items during online shopping. It extends the physical world try-on experience to the cyber space, enhancing the shopping experience for customers and thereby mitigating return and exchange rates, while concurrently bolstering sales and profits. Owing to its considerable value, virtual try-on has attracted growing interest from both academic and industrial community, leading to the emergence of numerous research works [14, 16, 21, 27, 42, 52].

Despite significant advancements, the task of virtual try-on has overlooked a crucial aspect within the realm of online shopping, that is mix-and-match. Initially, Han *et al.* define the task virtual try-on [18] as putting on a given fashion item to a partially masked query image. Hence, it only focuses on the quality of image generation regarding human pose, clothes deform and visual patterns, while overlooking a key question: whether the fashion item is a good match to the given clothes and person? In essence, this question corresponds to the task of fashion mix-and-match, a classical problem in both conventional fashion research and the recently fast-evolving computation fashion [11]. Especially in the community of multimedia and information retrieval, researchers have developed

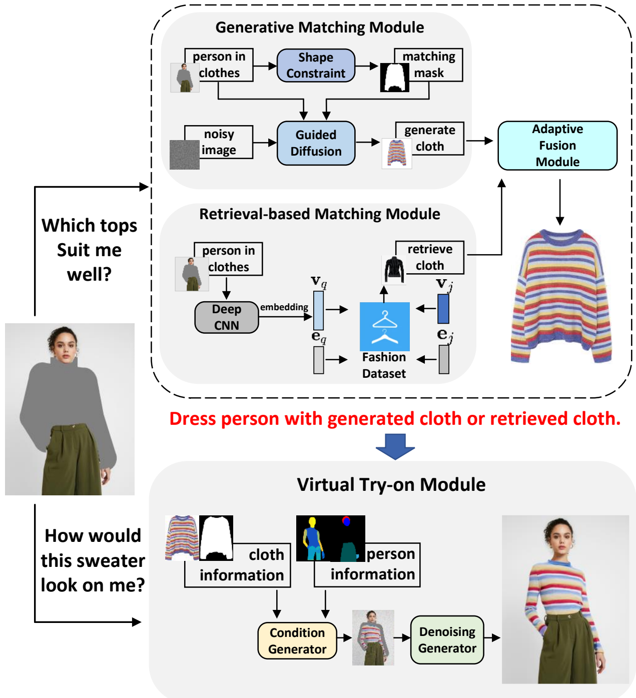

Figure 1: Mix-and-match and try-on are two essential fashion needs in daily life. We propose a one-stop system of Smart Fitting Room, which will generate or retrieve an apparel to match with the query and put it on the query image.

various algorithms, including both retrieval-based [6, 20, 31, 46] and generative methods [24, 25, 39], for fashion mix-and-match. Naturally, mix-and-match and virtual try-on are two essential fashion needs that are highly correlated with each other. Imagine a common shopping scenario: customers would first pick clothes, then try-on, and pick another. Such an interleaved iterative process repetitively plays everyday in online shopping, therefore, it motivates us that why don't we integrate these two separate tasks into a unified one-stop framework?

Following this motivation, in this paper, we seek to develop a novel system, *i.e.,* Smart Fitting Room, for the novel task of matching-aware virtual try-on, as shown in Figure 1. This system provides users with diverse, well-coordinated, and stylish clothing combinations, along with visual images of the virtual try-on effects. Compared with the conventional virtual try-on systems, our system can put on diverse and well-matched clothes to the given person. At the same time, comparing to the previous mix-and-match methods, our system is capable of demonstrating the synthesized try-on effects to the users, thus improving the consumers' shopping experience and the convert rate of the platforms. Despite various benefits, developing such a system is not simply chaining two separate models. Conversely, the distinctive motivation and characteristics of these two tasks pose exceptional challenges to the implementation of this novel system. An effective mix-and-match model often requires a large-scale dataset with tens of thousands of fashion items. It is not only necessary for training but also crucial for diversified recommendations since a limited number of options can hardly satisfy consumers' various fashion tastes. However, most of the online fashion stores can never access such a large-scale dataset. Moving one step back, even though they are able to procure such large-scale dataset in some way, it is still difficult for them to benefit from the external dataset. This is because existing mix-and-match models retrieve or generate items unrestricted, it is inevitable to recommend some items that are not served by itself or even non-existent, and recommending such items is commercially helpless to the pertinent store. In summary, there are long-standing yet overlooked challenges in developing a one-stop system for smart fitting room.

Addressing the above challenges, we propose Hybrid Matching-aware Virtual Try-On Framework (HMaVTON), encapsulating a hybrid mix-and-match module and an enhanced virtual try-on module. Specifically, we innovatively combine two mix-and-match models, which are separately trained following two distinctive paradigms, *i.e.,* one is retrieval-based and the other is generative. Upon two sets of recommended items, we design a simple yet effective adaptive fusion method to ground the generated items to the retrieved items. As a result, we can offer a hybrid list of fashion items, where the grounded items that exist on the platform can help boost sales and profits, while the generated items can satisfy consumers' diverse tastes and improve the shopping experience. To be noted, our fusion method is controllable with a simple threshold, therefore, we can smoothly change the ratio between generated and retrieved items. We integrate this hybrid mix-and-match module to an enhanced virtual try-on module, resulting in the final framework HMaVTON. In terms of the evaluation, since there is no available ground-truth for matching-aware try-on systems, we collaborate with fashion designers and conduct expert-level human evaluation. The experimental results demonstrate that our framework offers one-stop shopping service. In particular, the hybrid mix-and-match module can yield best matching score, and our enhanced virtual try-on module can generate try-on effects of higher quality. In summary, **the primary contributions** of our work are as follows:

- We introduce a novel task of matching-aware virtual try-on, which is the first time to integrate two essential fashion needs, *i.e.,* mix-and-match and virtual try-on, into a unified framework.
- We present the Hybrid Matching-aware Virtual Try-On Framework (HMaVTON), where the hybrid mix-and-match module adaptively fuses both retrieval-based and generative matching results, reaching a controllable balance between user experience and commercial benefits.
- We collaborate with fashion designers and conduct an expert-level human evaluation. And the results indicate that our framework can yield best performance.

2 RELATED WORK

Virtual Try-on: Virtual try-on [5, 10, 23] aims to seamlessly transfer clothing onto specific characters. Previous works [9, 28, 40, 42, 48] involves generating warped clothes aligned with character, and then generating images of the character wearing the warped clothes. In the clothes warping stage, VITON [18] introduce a process guided by TPS [4] for local non-rigid deformation of clothes. Subsequent research [1, 7, 17, 21] introduce flow-based methods. In the clothes and character synthesis stage, previous approaches [14, 18, 21, 27, 42] have predominantly relied on GANs, which are susceptible to mode

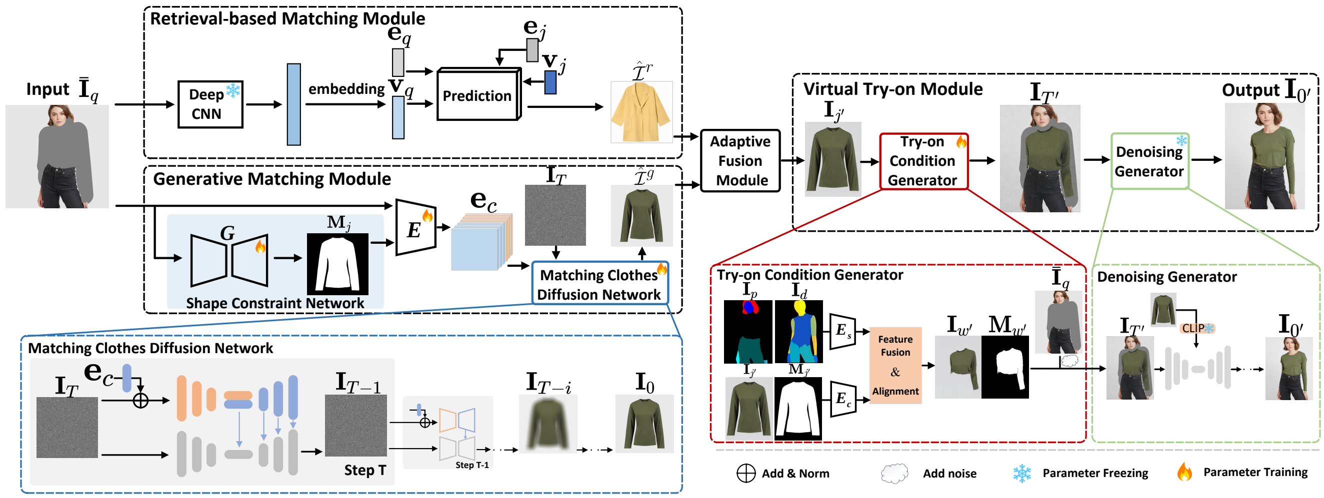

Figure 2: A schematic of our one-stop framework. We adaptively fuse retrieval-based and generative matching results using the Adaptive Fusion Module, and provide high-quality try-on effects through the Virtual Try-on Module.

collapse during training. Current research [16, 52] are increasingly adopting diffusion-based methods. But users are constrained by the clothes items provided in the dataset, making it challenging to validate the rationality of their outfit choices.

Fashion Mix-and-Match: Mix-and-match is a popular task in the domain of computational fashion [11], where the major solutions are based on item-based recommendation models for either pair-wise matching [12, 30, 44–46] or outfit recommendation [13, 32]. However, the scarcity of user interactions with clothes items complicates the use of standard recommender systems. In previous works, fashion mix-and-match clothes often utilizes visually-perceived collaborative filtering techniques. Approaches like VBPR [20] extract image category features and augment item vectors with them. But they are constrained by existing clothes datasets and can only recommend pre-existing clothes pairings. The generative fashion recommendation models [24–26, 39, 43, 47] have the capability to create novel fashion mix-and-match clothes, providing more creative and personalized fashion suggestions. Previous works, such as CRAFT [24], generate feature representations for clothes pairings and retrieve the most suitable individual clothes items from the dataset. DVBPR [25] generates clothes images based on user preferences but is limited to generating images that are identical in shape to those in the dataset. Despite these efforts, they are still constrained by the dataset and often produce unsatisfactory results. We introduce a novel diffusion model-based fashion recommendation model that not only produces high-quality and diverse clothes items but also generates visual images of users wearing clothes, significantly enhancing the user experience.

3 METHOD

We first present the problem formulation our proposed new task of matching-aware virtual try-on. Thereafter, we describe our model design, consisting of two key modules: 1) the mix-and-match module, which innovatively unifies both retrieval-based and generative methods for versatile fashion matching; and 2) the virtual try-on module, which optimizes the clothes warping effects to enhance the detail quality of the generated images.

3.1 Problem Formulation

We first revisit the problem formulations of mix-and-match and virtual try-on, followed by our proposed novel formulation of matching-aware virtual try-on.

Mix-and-Match. Given a query fashion item i and its associated image $\mathbf{I}_i \in \mathbb{R}^{3\times H\times W}$, where H, W are the height and width of the image, respectively, the task of mix-and-match aims to offer another fashion item j that matches well to i in terms of functionality, style, color, *etc.*There are two typical paradigms for mix-and-match: 1) retrieval-based methods, which aim to retrieve the top-k fashion items from a set of candidate items $\mathcal{I} = \{i_1, i_2, \cdots, i_N\}$, where N is the size of the item set and the ground-truth item $j \in \mathcal{I}$; and 2) generative methods, which aim to train a generation model that directly generates an item j, *aka.* the image of item j, where $j \in \mathcal{I}$.

Virtual Try-on. We have a ground-truth image $\mathbf{I}_q \in \mathbb{R}^{3\times H\times W}$ that depicts a person wearing a pair of clothes (*i.e.,* the top item j and the bottom item i). We mask the area of one item (*i.e.,* the top item j) in $\mathbf{I}_q$ and obtain the partially-masked image $\bar{\mathbf{I}}_q \in \mathbb{R}^{3\times H\times W}$. Given the partially-masked query image $\bar{\mathbf{I}}_q$ and the image $\mathbf{I}_j$ that is a product image of the masked fashion item j, the task of virtual try-on aims to restore the original $\mathbf{I}_q$, *i.e.,* trying this product j on the given person. To be noted, the try-on item $\mathbf{I}_j$ is provided as input, while the core of virtual try-on is make the synthesized try-on image convincing by considering the pose, deformation, and visual patterns [18].

Matching-aware Virtual Try-on. We propose a novel formulation to integrate the function of mix-and-match into the virtual try-on task. Specifically, given a partially-masked image $\bar{\mathbf{I}}_q$ (where the top item j is masked and the bottom item i is shown), this module needs to first recommend an item j' that is well matched to the presented item i, then put on its corresponding image $\mathbf{I}_{j'}$ to $\bar{\mathbf{I}}_q$, thus to recover a completely-dressed image $\mathbf{I}_q$. To be noted, the product item j' is not necessarily to be same with the originally-masked item j, just to make sure that j' is matched well with i conditioned on the person presented in $\bar{\mathbf{I}}_q$. In practice, the item j' can be retrieved from the candidate set $\mathcal{I}$ or generated from scratch.

3.2 Hybrid Mix-and-Match

Conventional approaches to fashion mix-and-match use either retrieval-based or generative methods, each of which has inherent limitations. Specifically, retrieval-based method can only recommend fashion items from an existing candidate pool, therefore, such method would fail if there is no suitable fashion items in stock on the given platform. In contrast, generative methods can address this problem owing to its capability in generating any fashion item without being limited to a given set of items. However, over reliance on generated items would harm the profits of the platform due to two reasons: 1) the existing fashion items will receive less exposure to users, undermining the sales of existing fashion items; and 2) the generated items require further customization and manufacture to convert to physical products, where the additional costs and longer leading time would depress or even lose the potential users.

To address the above limitations, we propose a novel hybrid mix-and-match method that takes advantage of both retrieval-based and generative methods, while preventing the pertinent drawbacks. Specifically, we first employ two types of methods to offer mix-and-match recommendations separately, then we propose an adaptive fusion strategy to combine both types of matching results.

3.2.1 Retrieval-based Matching Module. We follow one of the typical methods VBPR [20] to build the retrieval-based matching module. Specifically, given the partially-masked query image $\bar{\mathbf{I}}_q$ and its corresponding matched fashion items j, we employ the pre-trained deep CNN model, *i.e.,* ResNet-50 [19], to extract the visual features and then leverage a linear layer to transform the visual features into a shared representation space, which is formally defined as:

$$\begin{aligned} \mathbf{v}_q &= \text{ResNet}(\bar{\mathbf{I}}_q)\mathbf{W}_1, \\ \mathbf{v}_j &= \text{ResNet}(\mathbf{I}_j)\mathbf{W}_1, \end{aligned} \tag{1}$$

where $\mathbf{v}_q, \mathbf{v}_j \in \mathbb{R}^d$ are visual representations, $\mathbf{W}_1 \in \mathbb{R}^{2048\times d}$ is the feature transformation matrix of the linear layer, $\text{ResNet}(\cdot)$ represents the ResNet-50 network, and d is the dimensionality of the visual representation. Thereafter, we can calculate the matching score $\hat{x}_{q,j}$ via the following equation:

$$\hat{x}_{q,j} = \alpha + \beta_q + \beta_j + \mathbf{e}_q^\intercal \mathbf{e}_j + \mathbf{v}_q^\intercal \mathbf{v}_j, \tag{2}$$

where $\mathbf{e}_q, \mathbf{e}_j \in \mathbb{R}^d$ are the id embeddings of query image q and j, which are randomly initialized to pertain the collaborative filtering (CF) patterns [20]. α, β_q, β_j are trainable parameters to model the bias for the global, query q, and item j, respectively. We use Bayesian Personalized Ranking (BPR) [36] loss to train the model, denoted as:

$$\mathcal{L}^{\text{BPR}} = \sum_{(q,j,k)\in Q} -\ln\sigma(\hat{x}_{q,j} - \hat{x}_{q,k}), \tag{3}$$

where $Q = \{(q,j,k)|x_{q,j} = 1, x_{q,k} = 0\}$, $x_{q,j}$ is the ground-truth matching relation between query q and j, $x_{q,j} = 1$ indicates that (q,j) are matched with each other. In contrast, $x_{q,k} = 0$ means (q,k) is a an unmatched pair. $\sigma(\cdot)$ represents the sigmoid function. During inference, given a partially-masked query image, we rank all the items in the candidate set using the score function (defined in Equation 2) and take the top-k ranked items as the matched items, represented as $\hat{\mathcal{I}}^r = \{i_n^r\}_{n=1}^k$.

3.2.2 Generative Matching Module. The great success of Stable Diffusion [37] and its following work ControlNet [49] bring new opportunities for generative mix-and-match. Therefore, we propose a novel generative matching module based on ControlNet. Specifically, we first employ a Shape Constraint Network to generate the mask that depicts the shape of the desired fashion item image, then we use ControlNet the generated the matched image, by taking the generated mask as well as the original partially-masked query image as control conditions.

Mask Generation via GAN. We utilize a GAN [15] characterized with U-net [38] to generate the mask of the clothes to be generated, which is called Shape Constraint Network. Specifically, the generator G takes the partially-masked query image $\bar{\mathbf{I}}_q$ as input and generates the mask $\mathbf{M}_j$, which corresponds to the shape of the ground-truth product j. To be noted, the mask $\mathbf{M}_j$ is for the product image instead of the masked area in the query image. The discriminator D takes in $\mathbf{M}_j$ and discriminates if it a generated mask or the ground-truth mask. To optimized this network, we take advantage of both conditional GAN [33] and IoU [51], thus yielding a combined loss function. Owing to such a combination, empirical results show that it can generate high-quality masks. The loss function is defined as:

$$\mathcal{L}_G = \arg\min_G \max_D \mathcal{L}_{\text{cGAN}}(G,D) + \lambda\mathcal{L}_{\text{IoU}}(G). \tag{4}$$

Image Generation via ControlNet. We adopt ControlNet [49] to generate the final matching image. Concretely, we use encoder network E to convert pixel-space images (partially-masked query image $\bar{\mathbf{I}}_q$ and the generated mask $\mathbf{M}_j$) into latent images, then we concatenate both representations to form the control input of the ControlNet, formally represented as:

$$\mathbf{e}_c = \text{Concat}(E(\bar{\mathbf{I}}_q), E(\mathbf{M}_j)), \tag{5}$$

where $\mathbf{e}_c \in \mathbb{R}^{256\times 64\times 64}$, and $\text{Concat}(\cdot,\cdot)$ is the concatenate operation on the channel dimension. Thereafter, we leverage a typical ControlNet, and the predicted noise ϵ_θ is represented as following:

$$\epsilon_\theta = \text{F}(\mathbf{x};\Theta) + \text{Z}\left(\text{F}\left(\mathbf{x} + \text{Z}\left(\mathbf{e}_c;\Theta_z^1\right);\Theta_c\right);\Theta_z^2\right), \tag{6}$$

where $\mathbf{x}$ is the feature of noisy image $\mathbf{I}_t$ and $\text{F}(\cdot;\Theta_c)$ is the neural network blocks in the Stable Diffusion model [37], where Θ_c are the trainable parameters cloned from the original stable diffusion model. $\text{Z}(\cdot;\Theta_z)$ is a 1x1 convolution layer, where Θ_z^1 and Θ_z^2 are trainable parameters initialized with 0.

Given the text prompt p and the guiding condition vector $\mathbf{e}_c$, the algorithm learns a denoising network to predict the added noise. The learning objective is represented as:

$$\mathcal{L}_C = \mathbb{E}_{\mathbf{I}_0,t,p,e_c,\epsilon\sim\mathcal{N}(0,1)}\left[\|\epsilon - \epsilon_\theta\left(\mathbf{I}_t, t, p, \mathbf{e}_c\right)\|_2^2\right], \tag{7}$$

where time step t descends from T to 1 during the denoising process, and the input textual prompt p is simply set as "cloth". More sophisticated textual prompts are left for future study. Through the backward denoising process, we finally obtain the generated matching clothes $\mathbf{I}_{j'}$, *i.e.,* $\mathbf{I}_0$. Through running the generation process multiple rounds, we can also obtain a list of generated images, denoted as $\hat{\mathcal{I}}^g = \{i_n^g\}_{n=1}^k$.

3.2.3 Adaptive Fusion Module. After obtaining two sets of matching items $\hat{\mathcal{I}}^r$ and $\hat{\mathcal{I}}^g$, we propose an adaptive fusion module to combine both results for optimal outcome. The basic idea is to ground the generated images back to the retrieved images. Specifically, for every generated image $i_n^g \in \hat{\mathcal{I}}^g$, we calculate the distance between i_n^g and all the retrieved images $\hat{\mathcal{I}}^r$, where we use CLIP [35]

Figure 3: In comparison with baseline, our approach produces a diverse range of clothes styles, such as long sleeves, short sleeves, and straps, with color matching that adheres to the principles of harmony, creating a soothing visual resonance in terms of brightness and saturation.

to extract the image embeddings and cosine similarity as the distance metric. We set a threshold p, and if the cosine similarity is larger than p, we replace the generated image with the corresponding retrieved item. Therefore, the final results include both existing items and generated items, where the ratio can be adjusted by tuning the threshold p.

3.3 Virtual Try-on Module

When we generate the matched fashion item $i_{j'}$, our next step is to put on this generated image to the partially-masked query image $\bar{\mathbf{I}}_q$. We propose a new try-on model, which includes the Try-on Condition Generator and the Denoising Generator. The former generates the warped clothes $\mathbf{I}_{w'}$ and composites them onto the partially-masked query image $\bar{\mathbf{I}}_q$, while the latter applies noise addition and removal to the composited image, resulting in the visualized try-on image $\mathbf{I}_{0'}$.

3.3.1 Try-On Condition Generator. This module aims to obtain the warped clothes $\mathbf{I}_{w'}$ and corresponding mask $\mathbf{M}_{w'}$. It employs the concept of appearance flow [27] and consists of two encoders E_c and E_s. The clothes encoder E_c is used to extract features from clothes $\mathbf{I}_{j'} \in \mathbb{R}^{3\times h\times w}$ and mask $\mathbf{M}_{j'} \in \mathbb{R}^{h\times w}$. And the segmentation encoder E_s extracts features from the segmentation map $\mathbf{I}_d \in \mathbb{R}^{h\times w}$ and $\mathbf{I}_p \in \mathbb{R}^{3\times h\times w}$. The extracted features are fed into the flow pathway of the feature fusion block to generate the appearance flow map l_i. We obtain the warped cloth $\mathbf{I}_{w'}$ and its corresponding mask $\mathbf{M}_{w'}$ through feature fusion and alignment.

During the model training stage, the Try-on Condition Generator use $\mathcal{L}_{TV}$ to enhance the smoothness of the appearance flow:

$$\mathcal{L}_{TV} = ||\nabla l_4||_1. \tag{8}$$

And L1 loss and perceptual loss are used to encourage the network to deform clothes to fit the pose of the person:

$$\mathcal{L}_{L1} = \sum_{i=0}^{3} w_i \cdot \left\|\mathrm{D}\left(\mathbf{M}_{j'}, l_i\right) - \mathbf{M}_{w'}\right\|_1 + \left\|\hat{\mathbf{M}}_{w'} - \mathbf{M}_{w'}\right\|_1, \tag{9}$$

$$\mathcal{L}_{VGG} = \sum_{i=0}^{3} w_i \cdot \phi\left(\mathrm{D}\left(\mathbf{I}_{j'}, l_i\right), \mathbf{I}_{w'}\right) + \phi\left(\hat{\mathbf{I}}_{w'}, \mathbf{I}_{w'}\right), \tag{10}$$

where $\mathrm{D}(\cdot,\cdot)$ is defined as feature alignment, the operation involves removing non-overlapping regions. And w_i represents the weight of relative importance.

Table 1: Results of expert-level human evaluation. Our model has achieved the best performance in terms of overall user satisfaction and score, which indicates that the clothes recommended by our model not only meet the users' needs for matching but also offer them more diverse options.

Model	Index	Weight	Very Satisfied	Satisfied	Average	Dissatisfied	Very Dissatisfied	Score
GenMatching	D1-Style	24%	0.070	0.300	0.388	0.182	0.06	2.777
	D2-Color	25%	0.052	0.176	0.282	0.346	0.144	
	D3-Fabric	21%	0.054	0.224	0.344	0.296	0.082	
	D4-Variety	30%	0.048	0.118	0.346	0.292	0.196	
	Weighted score	-	0.056	0.198	**0.340**	0.280	0.126	
RetMatching	D1-Style	24%	0.116	0.41	0.346	0.110	0.018	3.489
	D2-Color	25%	0.120	0.376	0.330	0.152	0.022	
	D3-Fabric	21%	0.086	0.430	0.380	0.098	0.006	
	D4-Variety	30%	0.128	0.400	0.370	0.086	0.016	
	Weighted score	-	0.114	**0.403**	0.356	0.111	0.016	
GenMatching+	D1-Style	24%	0.100	0.418	0.364	0.112	0.006	3.420
	D2-Color	25%	0.090	0.348	0.346	0.200	0.016	
	D3-Fabric	21%	0.070	0.398	0.440	0.086	0.006	
	D4-Variety	30%	0.076	0.420	0.390	0.106	0.008	
	Weighted score	-	0.084	**0.397**	0.383	0.127	0.009	
RetMatching+	D1-Style	24%	0.122	0.456	0.314	0.090	0.018	3.557
	D2-Color	25%	0.118	0.370	0.346	0.128	0.038	
	D3-Fabric	21%	0.128	0.460	0.322	0.078	0.012	
	D4-Variety	30%	0.144	0.452	0.310	0.082	0.012	
	Weighted score	-	0.129	**0.434**	0.322	0.095	0.020	
HMaVTON(Ours)	D1-Style	24%	0.120	0.476	0.302	0.092	0.010	**3.578**
	D2-Color	25%	0.114	0.382	0.35	0.134	0.02	
	D3-Fabric	21%	0.132	0.472	0.314	0.074	0.008	
	D4-Variety	30%	0.13	0.464	0.312	0.086	0.008	
	Weighted score	-	0.124	**0.448**	0.32	0.097	0.012	

In summary, Try-on Condition Generator is trained using the following loss function, where λ_{L1}, and λ_{TV} represent hyperparameters controlling the relative importance between different losses:

$$\mathcal{L}_{CG} = \lambda_{L1}\mathcal{L}_{L1} + \mathcal{L}_{VGG} + \lambda_{TV}\mathcal{L}_{TV}. \tag{11}$$

The warped clothes $\mathbf{I}_{w'}$ and corresponding mask $\mathbf{M}_{w'}$ are combined with the partially-masked query image $\bar{\mathbf{I}}_q$, and randomly add t steps of Gaussian noise to obtain $\mathbf{I}_T^{'}$.

3.3.2 Denoising Generator. The noisy image $\mathbf{I}_T^{'}$ is denoised under the guidance of external condition $\mathbf{I}_{j'}$. The diffusion model loads parameters [16] as initial parameters. We encode clothes $\mathbf{I}_{j'}$ using the CLIP [35] image encoder to obtain the conditional input. This condition is injected into the network layers through cross-attention. Since the noisy image $I_T^{'}$ contains the original information of the clothes, the module achieves a more accurate representation of the wearing effect during the reverse diffusion process.

4 EXPERIMENTS

4.1 Experimental Settings

4.1.1 Datasets. We employ two datasets, *i.e.,* POG [7] and VITON-HD [8], where POG is used as an external dataset for the task of mix-and-match and VITON-HD is directly used for the evaluation of virtual try-on and the overall framework. We perform n-core filtering on the POG dataset, where we keep only the fashion items that occur between 5 and 100 times, resulting in 119,978 top-bottom pairs, 14,064 tops, and 8,124 bottoms. The VITON-HD dataset comprises 13,679 image pairs, consisting of frontal-view images of women and upper body images. To mitigate the domain gap between these two datasets, we perform an additional preprocessing the product images in the POG. Specifically, for the images of bottom clothes in POG dataset, we adjust their size and move them to the lower part of the overall image. Hence, the bottom images in POG would be better aligned with the partially-masked query images in the VITON-HD dataset. This allows the hybrid mix-and-match model to focus more on the matching relationship instead of the distribution gap regarding image size and position.

4.1.2 Evaluation Metrics. Our task aims to recommend a fashion item and try it on the body. Since the recommended items are not necessary to be identical to the item provided in dataset, how to evaluate the fashion matching and virtual try-on effects is challenging, especially for quantitative evaluation. To conduct a professional and faithful evaluation, we employ an expert-level human evaluation approach. Specifically, we invite a fashion researcher who is majoring in fashion design to propose a list of evaluation protocol and make a questionnaire. Thereafter, we make survey over ten fashion designers and obtain the final quantitative results.

The evaluation mainly consists of two parts: the rationality of clothes combinations and the diversity of clothes pairing effects. The former is typically analyzed in terms of clothes styles, colors, and fabrics. Specifically, the coordination of clothes styles should adhere to common pairing rules. Color coordination emphasizes the reasonable use and combination of hues. Fabric combinations should align with the season, in which the entire outfit is worn. The diversity primarily focuses on the mutual distinctions in the generated matching combinations, aiming to produce as many different coordination effects as possible.

To evaluate the virtual try-on performance, we employ the following quantitative analysis metrics: SSIM [41], FID [34], KID [3], IS [2], LPIPS [50], and PSNR [22]. Moreover, we use YOLOScore [29] and IS to justify the effectiveness of the SCN.

4.1.3 Compared Methods. Since the matching-aware virtual try-on task is newly proposed, there is no baseline or SOTA method for this task. To validate the effectiveness of our coordination, we design several baselines, which differ in various coordination modules.

Figure 4: Compared to other try-on models, our method exhibits superior performance in terms of clothes integrity restoration.

Table 2: Comparing to other virtual try-on methods, our model achieves the best performance across various metrics.

Method	SSIM ↑	FID ↓	KID ↓	IS ↑	LPIPS ↓	PSNR ↑
PF-AFN	0.8324	35.1156	3.0364	3.5732	0.1026	21.3806
FS-VTON	0.8335	35.1156	3.1431	3.5295	0.0913	21.9229
HR-VITON	0.8816	9.6404	0.2096	3.5305	0.0847	21.9228
GP-VTON	0.8874	9.2379	0.1683	3.5803	0.0713	23.0265
DCI-VTON	0.8952	8.3760	0.0292	3.6785	0.0698	23.9304
HMaVTON(Ours)	**0.8996**	**8.3349**	**0.0284**	**3.6787**	**0.0671**	**23.9782**

Specifically, we categorize the matching models into generation-based models and retrieval-based models. Additionally, to validate the effectiveness of our virtual try-on module, we conduct a comparative analysis of results with the latest try-on models.

Generative Models. The generation-based models create new clothes combination suggestions based on input images of users. The recommend item j' is generated from scratch by GMM. We define GenMatching is the baseline that does not incorporate external matching knowledge, while GenMatching+ understands.

Retrieval-based Models. The retrieval-based model incorporates visual information into recommendation system, returning the top k reasonable clothing combinations from the dataset. The recommend item j' is retrieved from the candidate set $\mathcal{I}$ by RMM. Similarly, we set the baselines for using external matching dataset (POG) RetMatching+ and without introducing it as RetMatching (that only uses the image pairs in the VITON-HD dataset).

Visual Try-on Model. We also compare the try-on performance with existing try-on models, including PF-AFN [14], FS-VTON [21], HR-VITON [27], GP-VTON [42], and DCI-VTON [16].

4.1.4 Implementation Details. Our model is trained on four NVIDIA RTX 3090 GPUs with the Pytorch framework. The Retrieval-based Matching Module and the Generative Matching Module are both trained on POG dataset and then further fine-tuned on the VITON-HD dataset. The Retrieval-based Matching Module is trained for 100 iterations and fine-tuned for 80 iterations. The Generative Matching Module is trained for 40 epochs and fine-tuned for 50 epochs.

4.2 Quantitative Results

Table 1 shows the results of our method and the matching model in human evaluation. We invite ten professional fashion designers aged 18-30 to conduct expert-level human evaluations. We use a

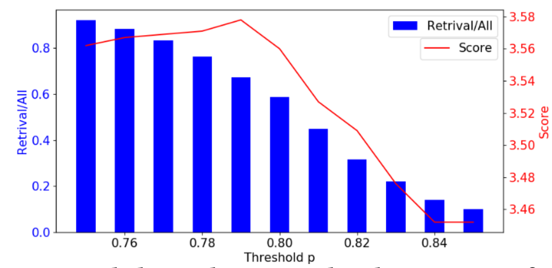

Figure 5: We balance the retrieval and generation of outfits by controlling the threshold to regulate the proportion of retrieved clothes in all the recommended clothes.

five-level scoring protocol to score 50 sets of results generated by 4 different models regarding to four aspects of D1-D4 (style, color, fabric, variety). The scores are discretized to five levels: very satisfied, satisfied, average, dissatisfied, and very dissatisfied, corresponding to 5 to 1 point, respectively. The survey is conducted through both online and offline manner. Finally, we collect valid scoring results from 10 designers for analysis.

We construct the rating matrix $\mathbf{R}$ based on the average values obtained from the questionnaire. We calculate the weighted score $\mathbf{B} = \mathbf{AR}$, where the weight matrix $\mathbf{A} = [0.24, 0.25, 0.21, 0.30]$. The weight matrix $\mathbf{A}$ is determined based on the expert's domain knowledge, where five experts first go through the generated results and then decide the weights of the four evaluation metrics D1-D4. The specific weights can be found in Supplementary Material **??**. We calculate the score based on the 5-point Likert scale ratings. Table 1 shows the results. Importantly, the scores in the "Weighted Score" row are the final representative results, the best value are highlighted in bold. Through the comparison of samples from the four groups, we can see that our model achieves the best performance in overall satisfaction and scores. By incorporating external matching dataset, it can effectively enhance the mix-and-match performance.

Table 2 shows the comparison between our model and try-on methods in terms of clothes fitting effects. Notably, the enhancements in the SSIM, FID, and KID metrics show significant improvements, indicating that our generated images closely resemble real images in terms of brightness, contrast, and structure. In the case of the IS metric, there is little variation among different methods,

Figure 6: Qualitative ablation study of shape constraint network and matching source.

suggesting that the Try-on results are controllable and can genuinely reflect the effect of clothes on the human body, indicating good model stability. Additionally, our model also exhibits slight improvements in the LPIPS and PSNR metrics, indicating an overall perceptual advantage in our images.

Figure 5 demonstrates the study of the controllable adaptive fusion module, where the threshold p is set to smoothly control the proportion between generated items and retrieved items. As the threshold increases, the proportion of retrieved images gradually decreases, while the proportion of generated images rises. Notably, the score first increases and then decreases, indicating the existence of an equilibrium point. At this point, the recommended clothes not only exhibit high image quality but also possess a reasonable matching effect while satisfying the users' demands for diversity.

4.3 Qualitative Results

From the left three columns of Figure 3, we can see that the images generated by our model exhibit better visual results. Not only is there a greater variety of clothes styles, but the fabrics in clothes are mostly soft and skin-friendly. The full-body back view images in the middle three columns demonstrate that our generation module can provide more options for clothes matching. For example, the generated short sleeves may not exist in the clothes database or may not have been recommended, but they still look quite good when worn, which can effectively help users expand their own style. In the right three columns, we can notice that the generated clothes can be mapped back to the retrieved items, and those with distinctive characteristics and novel styles can also be recommended.

Figure 4 shows that our model maintains better clothes details and offers a richer texture effect in the generated images. In the first row of images, PF-AFN [14] and FS-VTON [21] fail to preserve the overall clothes integrity, HR-VITON [27] exhibits a noticeable protrusion on the right side of the garment's hem, GP-VTON [42] lacks texture on the left joints and contains blurry regions on the right side, while DCI-VTON's [16] clothes neckline is different from the source attire. In the second row of images, the clothes generated by PF-AFN, FS-VTON, and DCI-VTON all exhibit significant differences from the original attire, HR-VITON and GP-VTON have clothes sleeves that do not conform to body, whereas our generated images perform well in terms of texture and overall appearance.

Table 3: Metrics evaluate the generated results from shape and quality, indicating that the Full Model performs the best.

Method	YOLOScore↓	IS↑
w/o Shape Constraint	0.1830	1.5136
w/o Matching Source	0.1524	2.6420
Full Model	**0.1417**	**2.9149**

4.4 Ablation Study

When setting the baseline in Section 4.1.3, we have validated the effectiveness of incorporating external matching information. Next, we verify the impact of Shape Constraint Network and Matching Source in Generative Matching Module.

w/o Shape Constraint. Regarding the Generative Matching Module, we design ablated models to validate the effectiveness of the Shape Constraint Network within it. As shown in the first row of Figure 6, the absence of shape constraints leads to incomplete garments and boundary blending in the generated clothes.

w/o Matching Source. To verify the matching capability of our model, we remove the input of human image I. From the second row in Figure 6, it can be observed that without the inclusion of matching source information, the generated clothes exhibit a monotonous texture and do not match well with bottoms.

Table 3 indicates that under shape constraints, our model presents higher image quality and is more in line with aesthetic design. As shown in Figure 6, the issues such as incomplete clothes generation are less likely to occur. In terms of the effectiveness of model coordination validation, even without the input of person images, the model can still generate clothes. However, there are cases of monotonous textures and colors, suggesting that the model generates clothes without a clear understanding of the desired style. In contrast, our method not only meets the matching requirements but also ensures high-quality clothes generation.

5 CONCLUSION AND FUTURE WORK

We presented a comprehensive generative fashion mix-and-match clothes framework, allowing users to try different clothes combinations under our styling advice. To assess the matching capability of our model, we collaborated with fashion design researchers to design evaluation metrics and invited ten fashion design experts to assess the model. The results indicated that our hybrid mix-and-match approach is well received by the majority of designers in terms of style combinations, color coordination, and other aspects.

In future work, we will enhance the Hybrid Matching-aware Virtual Try-On Framework (HMaVTON) to meet coordination needs of lower garments, shoes, and other clothes items. Simultaneously, we will extend the external matching dataset, considering not only clothes combinations but also incorporating users' profile information. And we will further explore the optimal combination of retrieval-based and generative methods. Furthermore, we will integrate various fashion domain knowledge, such as styles and aesthetics, into our system to enhance the overall system.

ACKNOWLEDGEMENT

This research/project is supported by NExT Research Center.

REFERENCES

[1] Shuai Bai, Huiling Zhou, Zhikang Li, Chang Zhou, and Hongxia Yang. 2022. Single Stage Virtual Try-On Via Deformable Attention Flows. In *ECCV (15) (Lecture Notes in Computer Science, Vol. 13675)*. Springer, 409–425.

[2] Shane T. Barratt and Rishi Sharma. 2018. A Note on the Inception Score. *CoRR* abs/1801.01973 (2018).

[3] Mikolaj Binkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. 2018. Demystifying MMD GANs. In *ICLR (Poster)*. OpenReview.net.

[4] Fred L. Bookstein. 1989. Principal Warps: Thin-Plate Splines and the Decomposition of Deformations. *IEEE Trans. Pattern Anal. Mach. Intell.* 11, 6 (1989), 567–585.

[5] Chieh-Yun Chen, Ling Lo, Pin-Jui Huang, Hong-Han Shuai, and Wen-Huang Cheng. 2021. FashionMirror: Co-attention Feature-remapping Virtual Try-on with Sequential Template Poses. In *ICCV*. IEEE, 13789–13798.

[6] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive Collaborative Filtering: Multimedia Recommendation with Item- and Component-Level Attention. In *SIGIR*. ACM, 335–344.

[7] Wen Chen, Pipei Huang, Jiaming Xu, Xin Guo, Cheng Guo, Fei Sun, Chao Li, Andreas Pfadler, Huan Zhao, and Binqiang Zhao. 2019. POG: Personalized Outfit Generation for Fashion Recommendation at Alibaba iFashion. In *KDD*. ACM, 2662–2670.

[8] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. 2021. VITON-HD: High-Resolution Virtual Try-On via Misalignment-Aware Normalization. In *CVPR*. Computer Vision Foundation / IEEE, 14131–14140.

[9] Ayush Chopra, Rishabh Jain, Mayur Hemani, and Balaji Krishnamurthy. 2021. ZFlow: Gated Appearance Flow-based Virtual Try-on with 3D Priors. In *ICCV*. IEEE, 5413–5422.

[10] Chien-Lung Chou, Chieh-Yun Chen, Chia-Wei Hsieh, Hong-Han Shuai, Jiaying Liu, and Wen-Huang Cheng. 2022. Template-Free Try-On Image Synthesis via Semantic-Guided Optimization. *IEEE Trans. Neural Networks Learn. Syst.* 33, 9 (2022), 4584–4597.

[11] Yujuan Ding, Zhihui Lai, P. Y. Mok, and Tat-Seng Chua. 2023. Computational Technologies for Fashion Recommendation: A Survey. *CoRR* abs/2306.03395 (2023).

[12] Yujuan Ding, P. Y. Mok, Yi Bin, Xun Yang, and Zhiyong Cheng. 2023. Modeling Multi-Relational Connectivity for Personalized Fashion Matching. In *ACM Multimedia*. ACM, 7047–7055.

[13] Yujuan Ding, P. Y. Mok, Yunshan Ma, and Yi Bin. 2023. Personalized fashion outfit generation with user coordination preference learning. *Inf. Process. Manag.* 60, 5 (2023), 103434.

[14] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. 2021. Parser-Free Virtual Try-On via Distilling Appearance Flows. In *CVPR*. Computer Vision Foundation / IEEE, 8485–8493.

[15] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *NIPS*. 2672–2680.

[16] Junhong Gou, Siyu Sun, Jianfu Zhang, Jianlou Si, Chen Qian, and Liqing Zhang. 2023. Taming the Power of Diffusion Models for High-Quality Virtual Try-On with Appearance Flow. In *ACM Multimedia*. ACM, 7599–7607.

[17] Xintong Han, Weilin Huang, Xiaojun Hu, and Matthew R. Scott. 2019. ClothFlow: A Flow-Based Model for Clothed Person Generation. In *ICCV*. IEEE, 10470–10479.

[18] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S. Davis. 2018. VITON: An Image-Based Virtual Try-On Network. In *CVPR*. Computer Vision Foundation / IEEE Computer Society, 7543–7552.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *CoRR* abs/1512.03385 (2015).

[20] Ruining He and Julian J. McAuley. 2016. VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback. In *AAAI*. AAAI Press, 144–150.

[21] Sen He, Yi-Zhe Song, and Tao Xiang. 2022. Style-Based Global Appearance Flow for Virtual Try-On. In *CVPR*. IEEE, 3460–3469.

[22] Alain Horé and Djemel Ziou. 2010. Image Quality Metrics: PSNR vs. SSIM. In *ICPR*. IEEE Computer Society, 2366–2369.

[23] Chia-Wei Hsieh, Chieh-Yun Chen, Chien-Lung Chou, Hong-Han Shuai, Jiaying Liu, and Wen-Huang Cheng. 2019. FashionOn: Semantic-guided Image-based Virtual Try-on with Detailed Human and Clothing Information. In *ACM Multimedia*. ACM, 275–283.

[24] Cong Phuoc Huynh, Arri Ciptadi, Ambrish Tyagi, and Amit Agrawal. 2018. CRAFT: Complementary Recommendations Using Adversarial Feature Transformer. *CoRR* abs/1804.10871 (2018).

[25] Wang-Cheng Kang, Chen Fang, Zhaowen Wang, and Julian J. McAuley. 2017. Visually-Aware Fashion Recommendation and Design with Generative Image Models. In *ICDM*. IEEE Computer Society, 207–216.

[26] Sudhir Kumar and Mithun Das Gupta. 2019. c^{+}GAN: Complementary Fashion Item Recommendation. *CoRR* abs/1906.05596 (2019).

[27] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. 2022. High-Resolution Virtual Try-On with Misalignment and Occlusion-Handled Conditions. In *ECCV (17) (Lecture Notes in Computer Science, Vol. 13677)*. Springer, 204–219.

[28] Kedan Li, Min Jin Chong, Jeffrey Zhang, and Jingen Liu. 2021. Toward Accurate and Realistic Outfits Visualization With Attention to Details. In *CVPR*. Computer Vision Foundation / IEEE, 15546–15555.

[29] Zejian Li, Jingyu Wu, Immanuel Koh, Yongchuan Tang, and Lingyun Sun. 2021. Image Synthesis from Layout with Locality-Aware Mask Adaption. In *ICCV*. IEEE, 13799–13808.

[30] Shuiying Liao, Yujuan Ding, and P. Y. Mok. 2023. Recommendation of Mix-and-Match Clothing by Modeling Indirect Personal Compatibility. In *ICMR*. ACM, 560–564.

[31] Qiang Liu, Shu Wu, and Liang Wang. 2017. DeepStyle: Learning User Preferences for Visual Recommendation. In *SIGIR*. ACM, 841–844.

[32] Yunshan Ma, Xiaohao Liu, Yinwei Wei, Zhulin Tao, Xiang Wang, and Tat-Seng Chua. 2024. Leveraging Multimodal Features and Item-level User Feedback for Bundle Construction. In *WSDM*. ACM, 510–519.

[33] Mehdi Mirza and Simon Osindero. 2014. Conditional Generative Adversarial Nets. *CoRR* abs/1411.1784 (2014).

[34] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. 2022. On Aliased Resizing and Surprising Subtleties in GAN Evaluation. In *CVPR*. IEEE, 11400–11410.

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML (Proceedings of Machine Learning Research, Vol. 139)*. PMLR, 8748–8763.

[36] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *UAI*. AUAI Press, 452–461.

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*. IEEE, 10674–10685.

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI (3) (Lecture Notes in Computer Science, Vol. 9351)*. Springer, 234–241.

[39] Yong-Siang Shih, Kai-Yueh Chang, Hsuan-Tien Lin, and Min Sun. 2018. Compatibility Family Learning for Item Recommendation and Generation. In *AAAI*. AAAI Press, 2403–2410.

[40] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. 2018. Toward Characteristic-Preserving Image-Based Virtual Try-On Network. In *ECCV (13) (Lecture Notes in Computer Science, Vol. 11217)*. Springer, 607–623.

[41] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* 13, 4 (2004), 600–612.

[42] Zhenyu Xie, Zaiyu Huang, Xin Dong, Fuwei Zhao, Haoye Dong, Xijin Zhang, Feida Zhu, and Xiaodan Liang. 2023. GP-VTON: Towards General Purpose Virtual Try-On via Collaborative Local-Flow Global-Parsing Learning. In *CVPR*. IEEE, 23550–23559.

[43] Yiyan Xu, Wenjie Wang, Fuli Feng, Yunshan Ma, Jizhi Zhang, and Xiangnan He. 2024. DiFashion: Towards Personalized Outfit Generation and Recommendation. *CoRR* abs/2402.17279 (2024).

[44] Xun Yang, Xiaoyu Du, and Meng Wang. 2020. Learning to Match on Graph for Fashion Compatibility Modeling. In *AAAI*. AAAI Press, 287–294.

[45] Xun Yang, Xiangnan He, Xiang Wang, Yunshan Ma, Fuli Feng, Meng Wang, and Tat-Seng Chua. 2019. Interpretable Fashion Matching with Rich Attributes. In *SIGIR*. ACM, 775–784.

[46] Xun Yang, Yunshan Ma, Lizi Liao, Meng Wang, and Tat-Seng Chua. 2019. TransNFCM: Translation-Based Neural Fashion Compatibility Modeling. In *AAAI*. AAAI Press, 403–410.

[47] Zilin Yang, Zhuo Su, Yang Yang, and Ge Lin. 2018. From recommendation to generation: A novel fashion clothing advising framework. In *2018 7th International Conference on Digital Home (ICDH)*. IEEE, 180–186.

[48] Ruiyun Yu, Xiaoqi Wang, and Xiaohui Xie. 2019. VTNFP: An Image-Based Virtual Try-On Network With Body and Clothing Feature Preservation. In *ICCV*. IEEE, 10510–10519.

[49] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *ICCV*. IEEE, 3813–3824.

[50] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*. Computer Vision Foundation / IEEE Computer Society, 586–595.

[51] Dingfu Zhou, Jin Fang, Xibin Song, Chenye Guan, Junbo Yin, Yuchao Dai, and Ruigang Yang. 2019. IoU Loss for 2D/3D Object Detection. In *3DV*. IEEE, 85–94.

[52] Luyang Zhu, Dawei Yang, Tyler Zhu, Fitsum Reda, William Chan, Chitwan Saharia, Mohammad Norouzi, and Ira Kemelmacher-Shlizerman. 2023. TryOnDiffusion: A Tale of Two UNets. In *CVPR*. IEEE, 4606–4615.