MMTryon: Multi-Modal Multi-Reference Control for High-Quality Fashion Generation

Xujie Zhang*
Sun Yat-Sen University
zhangxj59@mail2.sysu.edu.cn

Ente Lin*
Tsinghua University
linet22@mails.tsinghua.edu.cn

Xiu Li
Bytedance
lixiu.cv@bytedance.com

Yuxuan Luo
Bytedance
luoyuxuan@bytedance.com

Michael Kampffmeyer
UiT The Arctic University of Norway
michael.c.kampffmeyer@uit.no

Xin Dong
Bytedance
dongxin.1016@bytedance.com

Xiaodan Liang†
Sun Yat-Sen University
xdliang328@gmail.com

Abstract

This paper introduces MMTryon, a multi-modal multi-reference VIrtual Try-ON (VITON) framework, which can generate high-quality compositional try-on results by taking as inputs a text instruction and multiple garment images. Our MMTryon mainly addresses two problems overlooked in prior literature: 1) **Support of multiple try-on items and dressing style** Existing methods are commonly designed for single-item try-on tasks (e.g., upper/lower garments, dresses) and fall short on customizing dressing styles (e.g., zipped/unzipped, tuck-in/tuck-out, etc.) 2) **Segmentation Dependency**. They further heavily rely on category-specific segmentation models to identify the replacement regions, with segmentation errors directly leading to significant artifacts in the try-on results. For the first issue, our MMTryon introduces a novel multi-modality and multi-reference attention mechanism to combine the garment information from reference images and dressing-style information from text instructions. Besides, to remove the segmentation dependency, MMTryon uses a parsing-free garment encoder and leverages a novel scalable data generation pipeline to convert existing VITON datasets to a form that allows MMTryon to be trained without requiring any explicit segmentation. Extensive experiments on high-resolution benchmarks and in-the-wild test sets demonstrate MMTryon's superiority over existing SOTA methods both qualitatively and quantitatively. Besides, MMTryon's impressive performance on multi-items and style-controllable virtual try-on scenarios and its ability to try on any outfit in a large variety of scenarios from any source image, opens up a new avenue for future investigation in the fashion community. Project page is this link

*Equal contribution, interns at Bytedance
†Corresponding author

Preprint. Under review.

Figure 1: High-quality generation results of our MMTryon which is capable of following multi-modal multi-reference instructions to generate try-on results.

1 Introduction

VIrtual Try-ON (VITON) technology aims to generate photo-realistic images featuring individuals adorned in specific garments according to input images of both the garments and the person. Due to the promising potential to revolutionize the online shopping experience, VITON has garnered considerable attention both in the academic and industry communities.

Despite the substantial progress achieved in academic benchmarks [8, 10], previous methods [42, 14, 17, 4, 39, 28, 16] still fall short of meeting the standards required by real-world applications, especially when it comes to compositional try-on scenarios. In particular, they do not allow for: 1) **Multiple-garment Compositional Try-On**: Users should be able to freely choose and combine a diverse range of tops, bottoms, and accessories, rather than being confined to the single-garment try-on scenario. 2) **Flexible Try-On Styles**: Users should have the flexibility to select their preferred try-on style, allowing them to specify how a garment is worn, such as deciding whether a top should be tucked into the trousers or draped over them.

Due to their focus on single garments, these methods encounter challenges when multiple garments and accessories need to be changed simultaneously. Additionally, they struggle to model the subtle differences in how each piece of clothing is styled during a compositional try-on, such as the overlapping relationship between tops and bottoms. Furthermore, given a flat in-store image of a coat with a zipper, existing methods are limited to replicating the garment in its original state onto the target, without the ability to adjust for open or closed styles. This significantly restricts the freedom of the try-on experience.

Secondly, existing methods lack the capability to precisely identify try-on areas, often resorting to trained segmentation models [15, 26] for explicit try-on area localization. This reliance on segmentation models makes these approaches highly dependent on the model's precision, which can be unsatisfactory in complex scenarios, introducing additional noise during training. Additionally,

conditioning the generation on the segmented garment results in the loss of information regarding how the garment is worn, leading to highly unrealistic try-on outcomes, particularly for uniquely designed garments. Moreover, the dependence on segmentation models severely restricts the flexibility of the try-on process, as these models are typically confined to a predefined set of garment classes.

To address these challenges, we propose MMTryon, a more flexible try-on model capable of achieving multi-modal compositional dressing without the need for pre-segmenting the try-on areas. Firstly, we employ a novel scalable data generation pipeline to acquire an enhanced dataset. Then, based on the enhanced dataset, we propose a parsing free garment encoder that extracts necessary clothing information by leveraging cross-attention between try-on textual descriptions of areas of interest and clothing reference images. By circumventing the limitations imposed by the accuracy of segmentation models in existing VITON models, our model can more effectively handle the nuances of different clothing items and styles. Lastly, MMTryon utilizes a specially designed multi-reference image attention module and a textual cross-attention module to control the try-on of multiple garments. In particular, we employ a comprehensive try-on instruction for cross-attention, providing global information for each piece of clothing as well as information on how the garment is worn. Subsequently, the features obtained from the parsing free garment encoder are fused with the original image and passed through a multi-reference image attention module to infuse the information into the result image. This unique multi-modal compositional try-on training mode enables our model to visualize different garment combinations and styles in a more precise and flexible way.

Overall, our contributions can be summarized as follows: First, we propose MMTryon which, to our knowledge, is the first model to support the multi-modal compositional try-on task. MMTryon enables the combination of one or multiple garments for try-on, while also allowing the manipulation of try-on effects through textual commands. We further introduce a scalable data generation pipeline and a specially designed garment encoder. This eliminates the need for any prior segmentation networks during both the training and inference phases to get garments, relying solely on text, images of individual garment items, or model images to identify the areas of interest. Finally, extensive experiments conducted on public try-on benchmarks and in-the-wild test sets are performed, demonstrating that our approach surpasses existing state-of-the-art methods such as Outfit Anyone[3].Notably, our method achieves the most flexible and interactive dressing experience available to date.

2 Related work

GAN-Based Virtual Try-on. Traditional virtual try-on methods [36, 11, 44, 14, 17, 8, 42, 46, 41, 25, 4, 10, 40, 12, 21, 39] usually adopt a two-stage pipeline with Generative Adversarial Networks (GANs). The first stage employs an explicit warping module to deform in-shop clothing to the target shape, while the second stage uses a GAN-based generator to fuse the deformed clothing onto the target person. For theses methods, the synthetic quality largely depends on the quality of the deformation in the first stage, prompting current methods to emphasize enhancing the non-rigid deformation capabilities of the warping module.

Meanwhile, some GAN-based approaches [9, 40] have explored the compositional try-on task, utilizing a cyclical generation mode to achieve sequential try-on. However, when the outcome of compositional dressing is entirely determined by the try-on order, the lack of understanding of the semantic information of the garments often leads to unrealistic results. Furthermore, most GAN-based solutions mentioned above directly utilize UNet-based generators for try-on synthesis, with little exploration into how to enhance the generator's capabilities. This results in poor resolution and visual quality of the try-on outcomes, making it difficult to integrate them with diverse linguistic instructions.

Diffusion-Based Virtual Try-on. Compared to GAN-based models, diffusion models have made significant strides in high-fidelity conditional image generation [33, 34, 32]. Image-based virtual try-on is essentially a specialized form of the image editing/repair task conditioned on a given garment image. Therefore, a straightforward adaptation is to extend text-to-image diffusion models to accept images as conditions. While methods such as [43, 20, 7] have demonstrated their capabilities in virtual try-on, they fall short in preserving the texture details of the try-on results due to inadequate extraction of fine-grained features by the image encoder. Recent methods [28, 16] aim to integrate traditional GAN-based approaches with diffusion models. They employ explicit warping modules

[3]Outfit Anyone is available at `https://github.com/HumanAIGC/OutfitAnyone`

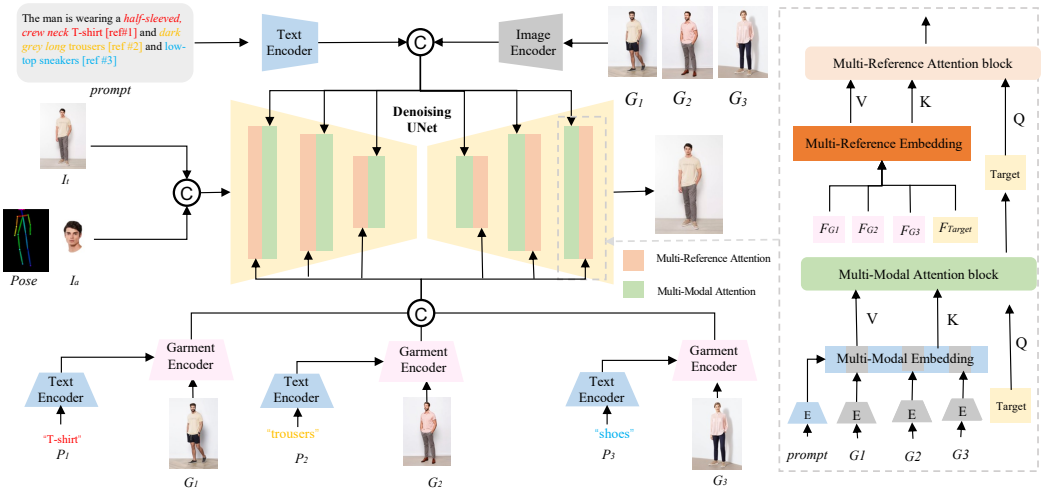


Figure 2: The framework of MMTryon. Clothing-agnostic representations, pose, and input noise are jointly fed into the model. The instruction prompt and garment images are combined to obtain a multi-modal embedding, conditioning the diffusion process on the overall image structure. Afterward, each garment image is individually processed through a garment encoder to obtain features F_{G1}, F_{G2}, F_{G3}, which, along with input features F_{target}, undergoes multi-reference attention for detailed texture transfer.

to create deformed garments and use diffusion models to blend these with reference person images. Beyond simply incorporating explicit warping modules, the newly introduced TryonDiffusion [47], conversely, implements an implicit warping mechanism for clothing deformation. This addresses the issues of texture misalignment mentioned earlier and achieves promising try-on synthesis. However, TryonDiffusion relies on precise parsing to achieve various conditions, making it unable to handle more complex try-on scenarios.

Additionally, none of the existing approaches have effectively leveraged the capabilities of current large multi-modal models and have discarded text as an input, lacking the ability of free-form instruction following. The try-on effect is solely inferred from the garment image.

Our approach, through the specialized design of a parsing free garment encoder and the multi-modal multi-reference attention mechanism, achieves a more versatile try-on experience. This not only addresses the limitations seen in both GAN and diffusion model-based methods but also utilizes the strengths of large multi-modal models, allowing for more detailed and flexible adaptation to various clothing styles and try-on scenarios.

3 Method

As shown in Fig. 2, MMTryon has a minimal input setup during inference, only requiring a source person image, multiple reference images, and a text instruction describing how items from references are fitted onto the desired person. To achieve multi-modal multi-reference controllable try-on, two key elements are central: the training data and the model design. We start by presenting the design of our model to support this inference setup, then we will describe how the data is gathered to support our model training.

3.1 MMTryon Model

3.1.1 Overview.

Given a target person image I, multiple reference images containing try-on items $\{G_i\}$ and a text instruction T describing how the items from those reference images are dressed, MMTryon aims to seamlessly follow T to synthesize photo-realistic try-on results I'. We model this task as a conditional diffusion model, which takes I, $\{G_i\}$ and T as conditions and gradually estimates a denoised version

of I' from pure noise. Our diffusion model is built upon the public available Stable Diffusion text-to-image model[1].

3.1.2 Stable Diffusion Model.

Stable Diffusion is a text conditioned latent diffusion model[33]. For an VAE [23] encoded image latent feature z_0, the forward diffusion process is performed by adding noise according to a predefined noise scheduler α_t[19]:

$$q(z_t|z_0) = \mathcal{N}(z_t; \sqrt{\alpha_t} z_0, (1-\alpha_t)I). \tag{1}$$

To reverse the diffusion process, an noise estimator $\epsilon_\theta(\cdot)$ parameterized by an UNet is learned to predict the forward added noise ϵ with the objective function,

$$\mathcal{L}_{\text{dm}} = \mathbb{E}_{(z_0,c)\sim D}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,1),t}\left[||\epsilon - \epsilon_\theta(z_t, t, c)||^2\right], \tag{2}$$

where c is the text condition associated with image latent z, and D is the training set. In the Stable Diffusion model, each block of the UNet is composed of cross-attention and self-attention layers. The cross-attention layer performs attention between the image feature query and the text condition embedding, and self-attention is performed within the image feature space.

Prior works [18, 13, 6] show that the cross-attention layer learns correspondences between the image feature and the text embedding tokens, generating an overall image semantic structure, while the self-attention layers are more related to detailed texture generation. Drawing inspiration from this, we extend Stable Diffusion by changing the mechanism of its attention layers to accept multi-modal multi-reference conditions. Specifically, we enhance the original text-conditioned cross-attention layer with our Multi-Modal Instruction Attention to generate the overall structure of the final try-on result. The self-attention layer is augmented with our Multi-Reference Texture Attention mechanism to seamlessly transfer detailed image features from garment images to the desired person image.

3.1.3 Multi-Modal Instruction Attention.

In our task, it is not possible to ask the user to describe the desired garments from the reference images precisely using only text, as high-level descriptions such as garment type exhibit ambiguity and can not guarantee dressing fidelity. To mitigate this issues, we form a multi-modal instruction using an interleaved prompt template:

A person wearing <garment><style> $[Ref\#1]$,...,and <garment><style>$[Ref\#N]$

For example, *A person wearing tops, tucked in, [Ref#1], pants [Ref#2] and shoes [Ref#3]*. Here *[Ref#N]* is a special placeholder token prepared for the corresponding reference image. This text prompt is first encoded using CLIP text encoders into text embedding $\mathbf{T} \in \mathbb{R}^{L\times D}$, where L is the length of the tokens and D is the embedding dimension. The embeddings corresponding to the placeholder tokens are then fused with the CLIP image embedding of the related garment images (see Fig. 2). To balance the embedding length and image feature granularity, and also better align the image feature with the text embedding, we do not only use the [CLS] token as in [43] , but instead use the latent features from the penultimate layers of the CLIP ViT encoder. This latent feature is further downsampled using a perceiver-resampler [3].

This multi-modal embedding then serves as the replacement of the original text embedding of the stable diffusion model and not only aids in the generation of dressing styles but also enhances the stability of the try-on, ensuring the correct mapping of the garment.

3.1.4 Multi-Reference Texture Attention.

The Multi-Modal Instruction Attention cannot fully preserve the details of the garment due to the use of the CLIP ViT image encoder. To guarantee the texture consistency, we re-use the same UNet structure (referred to as the garment encoder) as the the main diffusion UNet to extract more detailed features from the reference images.

In each self-attention layer of the diffusion UNet, we extract a set of garment features corresponding to that layer, denoted as $\{F_{G_1}, F_{G_2}, \ldots, F_{G_n}\}$. We denote the original feature map from the diffusion UNet as F_{target}. To perform the attention operation, the query Q is derived from F_{target}, while K, V pairs are obtained from the concatenated features of $\{F_{\text{target}}, F_{G_1}, F_{G_2}, \ldots, F_{G_n}\}$. This multi-reference warping attention allows us to transfer the features of different reference images onto

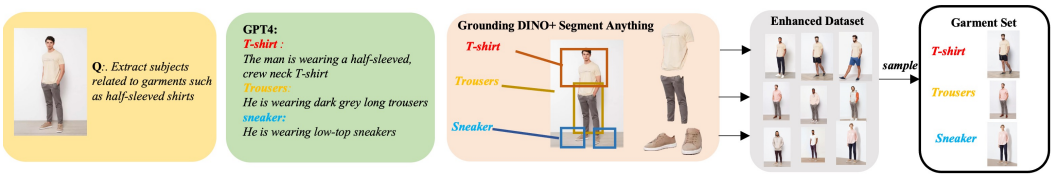


Figure 3: The data generation pipeline of MMTryon. We use a large multi-modal model to describe the target person image, followed by open-vocabulary grounding and segmentation models to extract correspondences between a person image and several garment subjects. For each subject, we utilize SDXL inpainting to obtain the enhanced dataset, which serves as our training data.

the target image, thus effectively preserving the texture features of each object without mutual interference. The process is illustrated in Fig. 2.

3.1.5 Parsing-Free Garment Encoder.

One of the main difference between our work and prior work, is that we do not rely on an off-the-shelf segmentation model. As mentioned in Sec. 3.1.2, the cross-attention layers of the Stable Diffusion UNet have shown potential to extract garment features corresponding to the text descriptions, secretly being a parsing model.

Thus for feature extraction, we prompt the garment encoder UNet with the textual label. As show in Fig. 2, the textual label P_i is derived from the instruction prompt T and used to extract features for the particular subject from image G_i, circumventing the use of an explicit segmentation model. These features of each attention layer are then injected into the main UNet to facilitate the multi-reference attention as mentioned above. Note that the pretrained diffusion UNet is not ready for extracting only the corresponding features, thus we will finetune this garment encoder jointly.

3.2 Scalable Data Generation Pipeline

To achieve multi-modal multi-reference try-on, a straightforward way is to construct training data in the same format. However, existing try-on datasets in this format are not readily available. There are generally two types of dataset, in-shop garment to person pairs or person to person pairs, where the former one only contains a single reference and is usually limited in garment types. Leveraging the person-to-person pairs instead, a possible way to automatically construct a training dataset for the multi-modal multi-reference try-on task is to use segmentation models to crop various reference garments to construct multi-reference pairs. However, the drawback of this approach is that the style of the garments is not controllable and that the model will still be dependent on off-the-shelf segmentation models during inference.

In this work, we instead propose to leverage the increasingly powerful capabilities of large models to develop a flexible scheme for dataset construction. Starting from existing person-to-person datasets, given a pair of person images I_a, I_b, we use SOTA large multi-modal model GPT-4V [2] and CogVLM [37] to caption the image in a desired format that is similar to the above-mentioned multi-modal instruction template, describing how the person is dressed. We then extract the main garment subjects and style (garment categories) from the image description denoted as $\langle obj_1, obj_2, \ldots, obj_n \rangle$. Subsequently, we use a SOTA open-vocabulary detection model [27] to generate the bounding boxes corresponding to each garment subjects within the image. The SAM [24] model is then used to generate precise mask annotations. After this caption-grounding-segmentation procedures, we now established a correspondence between a target person image I_a with description, and all its segmented garments from I_b. We find that this process performs well for the vast majority of clothing categories. Note that we need to establish correspondences between two different poses, otherwise this try-on task will degenerate to a copy-paste problem.

To enable parsing-free inference, starting from those garment images, we then use the inpainting ability of the Stable Diffusion model (we use SDXL [30] here) with a given prompt to randomly generate an image where a person is wearing such a garment. We use random seeds and random prompts to generate multiple samples, and discard samples with high similarity to prevent information leakage. This way, by iterating through all garment subjects, we are able to construction a training pair, where the target image is described using text, and the reference garments are from different

Figure 4: Qualitative comparisons on VITON-HD and DressCode in the single try-on task. Compared with other methods, our method MMTryon produces more realistic and texture-consistent images.

images. The overall process is illustrated in Fig. 3. To further make use of the garment-to-person pairs, we first train a single reference try-on network using a similar model design as described above, and transfer those images to person-to-person pairs, generating synthetic garment-person-person pairs. We now have constructed a multi-reference multi-modal instruction dataset that includes synthetic multi-reference data and the single-reference garment-person pairs.

4 Experiment

4.1 Datasets

We conduct extensive experiments on two public high-resolution Virtual Try-ON (VITON) benchmarks, namely VITON-HD [8] and DressCode [10], as well as on an additional large-scale proprietary e-commerce dataset. Experiments are conducted under a resolution of 1024×768. Specifically, VITON-HD comprises 13,679 image pairs of front-view upper-body women and upper-body in-store garment images, which are further divided into 11,647 training pairs and 2,032 testing pairs. DressCode includes 48,392 training pairs and 5,400 testing pairs of front-view full-body person and in-store garment images, consisting of three subsets with different category pairs (i.e., upper, lower, dresses). Our proprietary e-commerce dataset contains 57,239 training pairs and 5,219 testing pairs of front-view full-body person and in-store garment images, and 200,931 training pairs and 29,198 testing pairs of multiple front-view images of the same full body.

4.2 Implementation Details

During the data augmentation process, we enhanced the data across the three datasets, running the entire enhancement workflow on 32 A100 GPUs for 5 days. The augmented dataset comprises approximately 2 million images. In our MMTryon implementation, we utilize the SD-inpainting 2.0 pretrained model [33] to initialize the weights of our main UNet and employ SD 2.0 to initialize the weights of our auto parsing encoder. For the encoding of both text and global image information, we use clip-large-14 [31]. The entire model was trained on 8 A100 GPUs for 6 days with a batch size of 4 and a learning rate of $1e-5$.

4.3 Baselines

In our study, we initially compare our approach on traditional try-on tasks with previous baseline methods. We selected four advanced GAN-based methods, namely PF-AFN [14], FS-VTON [17], SDAFN [4], GP-VTON [39] and the latest diffusion-based methods, DCI-VTON [16] and LADI-VTON [28]. We directly utilize their released pretrained models for this comparison. Unfortunately, we are not able to conduct an extensive comparison with TryonDiffusion [47] as it is not publicly available at the time of submission. Secondly, for the compositional and multi-modal try-on tasks, we compare our work with paint by example [43], Stable diffusion [33], and DALLE3 [5]. For all baselines, we strictly adhere to the official instructions for running their training and testing

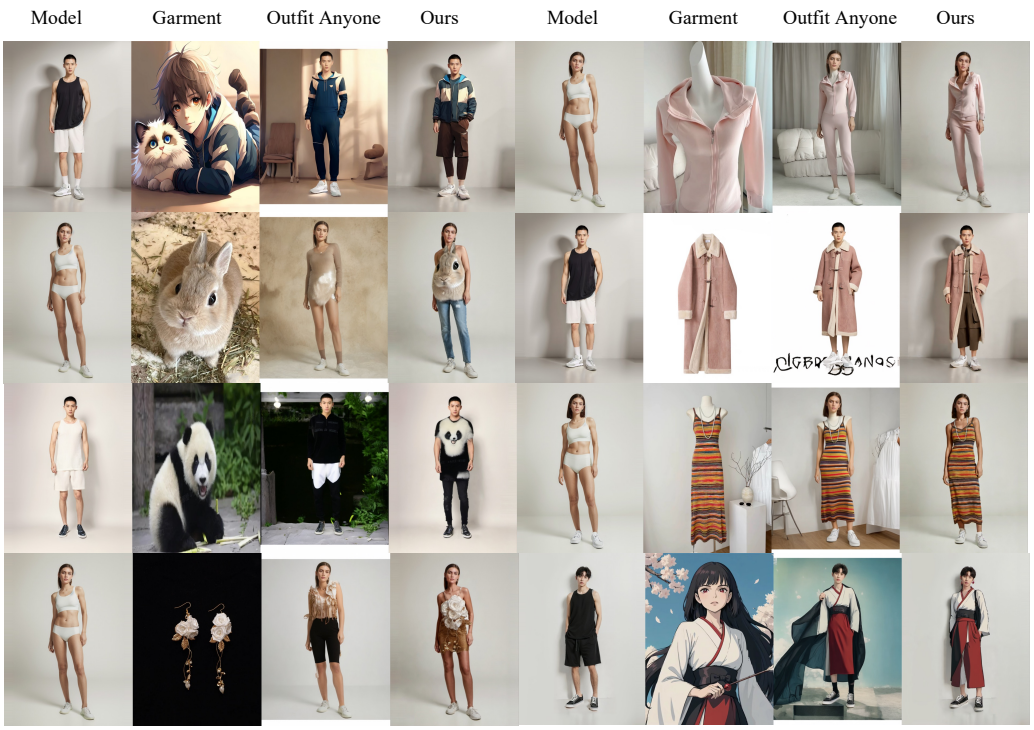


Figure 5: Qualitative comparisons in the wild. Compared with OutfitAnyone, our method MMTryon produces more realistic and stable images.Note, as Outfit Anyone is only available through their user interface, the comparisons here are limited to their provided model images.

scripts. Additionally, to further demonstrate our models zero-shot capabilities, we further perform a comparisons with the state-of-the-art community model, Outfit Anyone[4]

4.4 Qualitative Results

Comparison on single garment try-on. Fig. 4 and Fig. 5 provides a qualitative comparison between MMTryon and the state-of-the-art baselines in the single garment try-on task on the VITON-HD dataset [8]. The results demonstrate the superiority of our MMTryon approach over the baselines. Firstly, our method employs a more powerful cloth encoder, which results in better texture consistency compared to the baseline. Secondly, as our approach utilizes more textual information, it achieves greater stability in try-on styles than the baseline, showcasing the effectiveness of our method.

Comparison on multi-modal and compositional try-on. Fig. 6 provides a qualitative comparison between MMTryon and the state-of-the-art baselines for the compositional try-on and multi-modal try-on tasks in the wild. The results demonstrate the superiority of our approach relative to the baselines. Firstly, for compositional try-on, our method exhibits better texture consistency and more realistic try-on effects compared to the baselines. Secondly, for multi-modal try-on, our approach shows a more refined performance in texture consistency. At the same time, baselines often overlook the textual descriptions of dressing styles, whereas our method demonstrates stronger text-image consistency, effectively reflecting the intentions expressed in the text. This demonstrates that our approach achieves state-of-the-art results for the compositional and multi-modal try-on tasks. In Fig. 7, we further showcase MMTryon's ability to generate results that take into account dressing style specifications.

4.5 Quantitative Results

Metrics. For traditional single-garment try-on tasks, we utilize four widely-used metrics, namely the Structural Similarity Index (SSIM)[38], Perceptual Distance (LPIPS)[45], Kernel Inception Distance

[4]Outfit Anyone is available at `https://github.com/HumanAIGC/OutfitAnyone`.

Figure 6: Qualitative comparisons with Paint-by-Example[43], Midjourney and DALL·E3[5] in the Multi-Modal Multi-Reference task. Compared with other methods, our method can achieve the best results both in terms of faithfulness and realism.

Figure 7: More results of MMTryon. Our method demonstrates high quality result in a large variety of scenarios and precise control over dressing styles through instruction-based manipulation.

[35], and Fréchet Inception Distance (FID) [29], to evaluate the similarity between synthesized and real images. For the compositional try-on task, we similarly employ FID and KID to measure the quality of image generation. In the context of the multi-modal instruction try-on task, we assess the quality of image generation using FID, while the Clip score is utilized to evaluate the consistency between instructions and the final generated images. Additionally, we conducted a Human Evaluation (HE) study, inviting 100 reviewers to assess the synthesis quality and text-image consistency across three scenarios: single-garment try-on, compositional try-on, and multi-modal instruction try-on

Table 1: Quantitative comparisons on the VITON-HD [8] and Dresscode [10].

Dataset	VITON-HD [8]				Dresscode [10]			
Method	SSIM ↑	FID ↓	LPIPS ↓	KID ↓	SSIM ↑	FID ↓	LPIPS ↓	KID ↓
PF-AFN [14]	0.885	9.616	0.087	3.85	0.898	9.807	0.096	4.52
FS-VTON [17]	0.881	9.735	0.091	3.69	0.896	9.667	0.097	4.85
SDAFN [4]	0.881	9.497	0.092	2.73	0.892	9.783	0.108	3.91
GP-VTON [39]	0.893	9.405	0.079	0.88	0.898	9.238	0.091	1.88
DCI-VTON [16]	0.868	9.166	0.096	1.10	-	-	-	-
StableVTON [22]	0.866	8.992	0.079	1.03	-	-	-	-
MMTryon	**0.912**	**8.702**	**0.069**	**0.58**	**0.941**	**8.127**	**0.089**	**0.43**

Table 2: Quantitative comparisons on the compositional garment try-on task.

Task	Compositional Try-on		Multi-Modal Try-on		
Method	FID ↓	KID ↓	FID ↓	KID ↓	CLIP Score ↑
Paint-by-Example [43]	12.429	4.89	15.231	5.21	0.61
Midjourney	9.231	3.41	10.243	3.45	0.78
DALL·E3 [5]	9.441	3.19	12.241	4.12	0.74
MMTryon (ours)	**8.902**	**0.47**	**8.187**	**0.42**	**0.91**

Table 3: Ablation results demonstrate the benefit of the proposed data augmentation pipeline and multi-reference attention.

Method	FID ↓	KID ↓	SSIM ↑	LPIPS ↓	CLIP Score ↑
MMTryon w/o augmented dataset	9.896	1.41	0.868	0.093	0.71
MMTryon w/o multi-reference attention	12.89	0.871	0.883	0.091	0.77
MMTryon (ours)	**8.902**	**0.58**	**0.912**	**0.069**	**0.93**

Evaluation. As reported in Tab. 1 and Tab. 2, our MMTryon consistently outperforms the baseline methods across all metrics, proving that MMTryon can generate try-on results with better visual quality in single garment try-on and compositional garment try-on tasks. In addition, for the multi-modal garment try-on tasks, the FID and KID scores reflect our method's superior generation quality, while the significant lead in CLIP scores indicates that our method faithfully reflects the information described in the text. Furthermore, we designed a human evaluation to compare the generated results of the baselines to our model. Higher human evaluation scores indicate that a larger proportion of participants prefer the outcomes of the given method. As shown in Fig. 8, the human evaluators prefer the results generated by MMTryon in most cases, indicating superior texture consistency and text-image consistency.

4.6 Ablation study

To validate the effectiveness of the proposed data pipeline for auto parsing, we conducted an ablation study by comparing MMTryon with a version not using the augmented dataset. Fig. 9(a) shows that the MMTryon model without the augmented dataset loses the ability to segment using text, leading to incorrect try-on representations. Tab. 3 also indicates that the complete MMTryon has higher generation quality and text-image consistency. This demonstrates that the generated try-on results are more accurate, and the generated images are more in line with the real distribution of the dataset, thereby validating that the data pipeline enhances the ability to auto parse using text.

Next, to verify the effectiveness of the parsing-free garment encoder, we conducted another ablation study. In this study, we compared MMTryon with a version that only utilizes the CLIP features of the clothing, where we replaced the multi-reference attention with the original self-attention module. We again compared the main metric scores and image quality and observed a decline in both image quality and texture consistency when removed. Meanwhile, Fig. 9(b) indicates that rich clothing features are key to maintaining texture consistency.

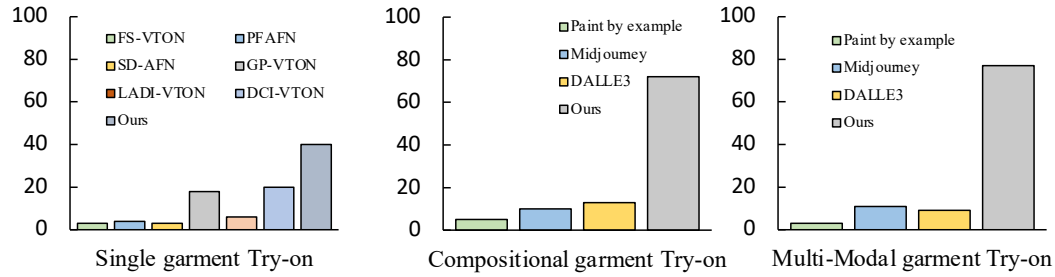
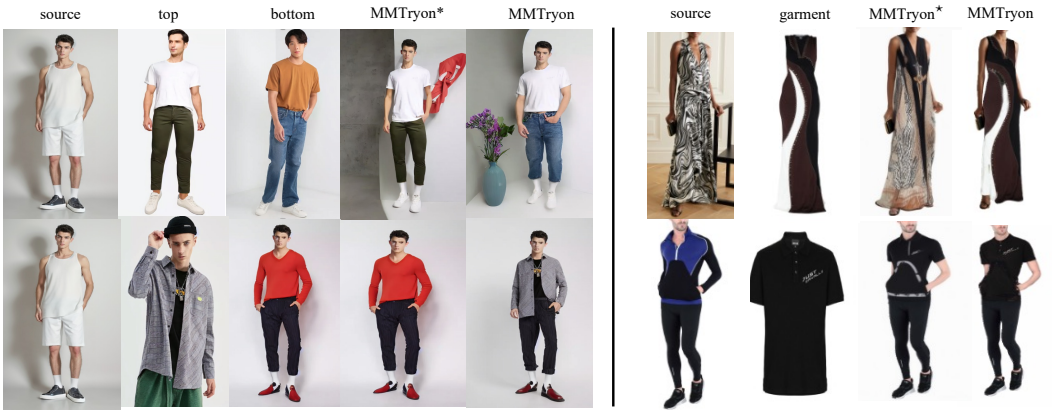


Figure 8: The results of the Human Evaluation (HE) study. Our proposed method outperforms the baseline approaches on all three tasks.



(a) Comparison with MMTryon without the augmented dataset

(b) Comparison with MMTryon w\o the multi-reference attention

Figure 9: Qualitative ablation study highlighting the value of our data augmentation pipeline and multi-reference attention.

5 Conclusion

In this work, we introduce MMTryon, a novel and powerful try-on model capable of freely generating high-fidelity VITON results with realistic try-on effects based on text and multiple garments. To eliminate the dependency on inefficient segmentation networks, MMTryon constructs a flexible and universal labeling model while redesigning the parsing-free garment encoder. To support multi-modal and multi-reference dressing modes, MMTryon introduces multi-modal instruction attention and multi-reference attention modules. Experiments conducted on high-resolution VITON benchmarks and various in-the-wild test sets have demonstrated MMTryon's superior efficacy compared to existing methods.

Societal impact. As with most generative approaches, MMTryon can be used for malicious purposes by generating images that infringe upon copyrights and/or privacy. Given these considerations, responsible use of the model is advocated.

Limitation and future work. While our method demonstrates strong performance, it still has certain limitations. In the data generation process, our method is influenced by the limitations of pretrained models, making it challenging to produce data that meets the requirements for very fine parts, such as cuffs and collars. This restricts our ability to generate detailed components. Moving forward, we may focus on fine-tuning large models to construct a more freely detailed and fine-grained dataset, aiming to enhance the upper limit of our model.

References

[1] Stable diffusion 2.0 inpainting. https://huggingface.co/stabilityai/stable-diffusion-2-inpainting

[2] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)

[3] Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022)

[4] Bai, S., Zhou, H., Li, Z., Zhou, C., Yang, H.: Single stage virtual try-on via deformable attention flows. In: European Conference on Computer Vision (2022)

[5] Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf (2023)

[6] Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. arXiv preprint arXiv:2304.08465 (2023)

[7] Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. arXiv preprint arXiv:2307.09481 (2023)

[8] Choi, S., Park, S., Lee, M., Choo, J.: Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2021)

[9] Cui, A., McKee, D., Lazebnik, S.: Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing. In: Proceedings of the IEEE/CVF international conference on computer vision (2021)

[10] Davide, M., Matteo, F., Marcella, C., Federico, L., Fabio, C., Rita, C.: Dress code: High-resolution multi-category virtual try-on. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)

[11] Dong, H., Liang, X., Shen, X., Wang, B., Lai, H., Zhu, J., Hu, Z., Yin, J.: Towards multi-pose guided virtual try-on network. In: Proceedings of the IEEE/CVF international conference on computer vision (2019)

[12] Dong, X., Zhao, F., Xie, Z., Zhang, X., Du, D.K., Zheng, M., Long, X., Liang, X., Yang, J.: Dressing in the wild by watching dance videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)

[13] Ge, S., Park, T., Zhu, J.Y., Huang, J.B.: Expressive text-to-image generation with rich text. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)

[14] Ge, Y., Song, Y., Zhang, R., Ge, C., Liu, W., Luo, P.: Parser-free virtual try-on via distilling appearance flows. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2021)

[15] Gong, K., Gao, Y., Liang, X., Shen, X., Wang, M., Lin, L.: Graphonomy: Universal human parsing via graph transfer learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)

[16] Gou, J., Sun, S., Zhang, J., Si, J., Qian, C., Zhang, L.: Taming the power of diffusion models for high-quality virtual try-on with appearance flow. In: Proceedings of the 31st ACM International Conference on Multimedia (2023)

[17] He, S., Song, Y.Z., Xiang, T.: Style-based global appearance flow for virtual try-on. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)

[18] Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)

[19] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems (2020)

[20] Huang, L., Chen, D., Liu, Y., Shen, Y., Zhao, D., Zhou, J.: Composer: Creative and controllable image synthesis with composable conditions. arXiv preprint arXiv:2302.09778 (2023)

[21] Huang, Z., Li, H., Xie, Z., Kampffmeyer, M., Liang, X., et al.: Towards hard-pose virtual try-on via 3d-aware global correspondence learning. Advances in Neural Information Processing Systems (2022)

[22] Kim, J., Gu, G., Park, M., Park, S., Choo, J.: Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)

[23] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)

[24] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)

[25] Lee, S., Gu, G., Park, S., Choi, S., Choo, J.: High-resolution virtual try-on with misalignment and occlusion-handled conditions. In: European Conference on Computer Vision (2022)

[26] Li, P., Xu, Y., Wei, Y., Yang, Y.: Self-correction for human parsing. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)

[27] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)

[28] Morelli, D., Baldrati, A., Cartella, G., Cornia, M., Bertini, M., Cucchiara, R.: LaDI-VTON: Latent Diffusion Textual-Inversion Enhanced Virtual Try-On. In: Proceedings of the ACM International Conference on Multimedia (2023)

[29] Parmar, G., Zhang, R., Zhu, J.Y.: On aliased resizing and surprising subtleties in gan evaluation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)

[30] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)

[31] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning (2021)

[32] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)

[33] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2022)

[34] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems (2022)

[35] Sutherland, J., Arbel, M., Gretton, A.: Demystifying mmd gans. In: International Conference for Learning Representations (2018)

[36] Wang, B., Zheng, H., Liang, X., Chen, Y., Lin, L., Yang, M.: Toward characteristic-preserving image-based virtual try-on network. In: Proceedings of the European conference on computer vision (ECCV) (2018)

[37] Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., et al.: Cogvlm: Visual expert for pretrained language models. arXiv preprint arXiv:2311.03079 (2023)

[38] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing (2004)

[39] Xie, Z., Huang, Z., Dong, X., Zhao, F., Dong, H., Zhang, X., Zhu, F., Liang, X.: 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 (2023)

[40] Xie, Z., Huang, Z., Zhao, F., Dong, H., Kampffmeyer, M., Dong, X., Zhu, F., Liang, X.: Pasta-gan++: A versatile framework for high-resolution unpaired virtual try-on. arXiv preprint arXiv:2207.13475 (2022)

[41] Xie, Z., Huang, Z., Zhao, F., Dong, H., Kampffmeyer, M., Liang, X.: Towards scalable unpaired virtual try-on via patch-routed spatially-adaptive gan. Advances in Neural Information Processing Systems (2021)

[42] Xie, Z., Zhang, X., Zhao, F., Dong, H., Kampffmeyer, M.C., Yan, H., Liang, X.: Was-vton: Warping architecture search for virtual try-on network. In: Proceedings of the 29th ACM International Conference on Multimedia (2021)

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

[44] Yang, H., Zhang, R., Guo, X., Liu, W., Zuo, W., Luo, P.: 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 (2020)

[45] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2018)

[46] Zhao, F., Xie, Z., Kampffmeyer, M., Dong, H., Han, S., Zheng, T., Zhang, T., Liang, X.: M3d-vton: A monocular-to-3d virtual try-on network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)

[47] Zhu, L., Yang, D., Zhu, T., Reda, F., Chan, W., Saharia, C., Norouzi, M., Kemelmacher-Shlizerman, I.: Tryondiffusion: A tale of two unets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)