

DiT-VTON: Diffusion Transformer Framework for Unified Multi-Category Virtual Try-On and Virtual Try-All with Integrated Image Editing

Qi Li^{1*} Shuwen Qiu^{2*†} Julien Han¹ Xingzi Xu^{1,3†} Mehmet Saygin Seyfioglu¹ Kee Kiat Koo¹
Karim Bouyarmane¹

¹Amazon ²University of California, Los Angeles (UCLA) ³Duke University

{qlimz, hameng, xingzixu, mseyfiog, kiatkoo, bouykari}@amazon.com

xingzi.xu@duke.edu

jantqiu@cs.ucla.edu

<https://dit-vton.github.io/DiT-VTON/>

Work submitted in November 2024 to CVPR 2025

Abstract

The rapid growth of e-commerce has intensified the demand for Virtual Try-On (VTO) technologies, enabling customers to realistically visualize products overlaid on their own images. Despite recent advances, existing VTO models face challenges with fine-grained detail preservation, robustness to real-world imagery, efficient sampling, image editing capabilities, and generalization across diverse product categories. In this paper, we present DiT-VTON, a novel VTO framework that leverages an architecture based on a Diffusion Transformer (DiT), renowned for its performance on text-conditioned image generation (text-to-image), adapted here for the image-conditioned VTO task. We systematically explore multiple DiT configurations, including in-context token concatenation, channel concatenation, and ControlNet integration, to determine the best setup for VTO image conditioning. To enhance robustness, we train the model on an expanded dataset encompassing varied backgrounds, unstructured references, and non-garment categories, demonstrating the benefits of data scaling for VTO adaptability. DiT-VTON also redefines the VTO task beyond garment try-on, offering a versatile Virtual Try-All (VTA) solution capable of handling a wide range of product categories and supporting advanced image editing functionalities, such as pose preservation, precise localized region editing and refinement, texture transfer and object-level customization. Experimental results show that our model surpasses state-of-the-art methods on public benchmark tests like VITON-HD, achieving superior detail preserva-

tion and robustness without reliance on additional image condition encoders. It also surpasses state-of-the-art models that have VTA and image editing capabilities on a varied dataset composed of thousands of product categories. As a result, DiT-VTON significantly advances VTO applicability in diverse real-world scenarios, enhancing both the realism and personalization of online shopping experiences.

1. Introduction

The rapid expansion of e-commerce has fueled demand for advanced tools that empower customers to virtually “try on” products by overlaying them onto their own images, enhancing both the personalization and interactivity of online shopping experiences. This technology, known as Virtual Try-On (VTO), requires a user’s image (source image), a specified mask indicating the intended placement of the product, and a product image (reference image). These components work together to seamlessly integrate the product into the designated area of the source image, creating a cohesive and realistic representation of the product on the user, as depicted in Fig. 1. From a technical perspective, the VTO task can be formulated as an image-conditioned inpainting problem, wherein the challenge lies in filling masked regions in the source image with highly realistic, contextually accurate product visualizations.

Recent breakthroughs in U-Net-based Latent Diffusion Models (LDMs) [1–4] have propelled the capabilities of VTO models by enabling them to leverage pre-trained weights from text-to-image and inpainting diffusion frameworks. These advancements have led to state-of-the-art (SOTA) performance, marked by superior rendering quality

*Equal contribution.

†Work done during internship at Amazon.



Figure 1. **Our DiT-VTON model applied to diverse use cases, enabling inpainting within user-specified editing regions using content guided by a reference image.** The model can semantically infer and generate expected objects, texture details, and even identify specific body parts for virtual try-on (VTO) tasks, showcasing its versatility in content-aware editing and synthesis. The model can also generate better product details and layouts for virtual try-all (VTA) tasks such as furniture. Please zoom in to see details preservation.

and realism, broadening the potential for widespread application in the online retail space.

Despite substantial progress, existing SOTA VTO models continue to encounter several key challenges that hinder their practical efficacy and versatility: **1. Fine-Grained Detail Preservation.** Maintaining fine-grained details (shown in Fig. 1a), such as logos, text decorations, and intricate textures, remains a challenge. These small yet crucial elements occupy only a minor portion of the generated image area, often leading models trained with denoising score matching objectives to overlook them. Preserving these details is essential for delivering visually accurate representations of products. **2. Robustness to In-the-Wild Images.** Current models are predominantly trained on preprocessed datasets featuring front-facing images of the person and the product, often set against a clean background. Consequently, performance deteriorates when faced with unstructured real-world conditions, such as cluttered backgrounds, people wearing garment or less organized garments in reference images (shown in Fig. 1c and Fig. 1e). Robustness across diverse, in-the-wild scenarios is critical for wider adoption and usability. **3. Simple and Efficient Sampling Process.** Integrating reference images and pose conditions requires external image encoders that embed image conditions into the generation space. ControlNet [5] based mod-

els [6, 7] are commonly used to facilitate the fusion of these images within the UNet framework. However, these additional modules increase computational overhead during both training and inference, impacting model efficiency and scalability. In addition, existing models rely on DDIM [8] for image sampling, which generally requires longer time steps.

In addition to these challenges, we believe that the scope of VTO should evolve beyond basic “garment swapping” to offer a more versatile, personalized virtual try-on experience across various product categories. This brings the following additional challenges: **4. Image Editing and Customization Capabilities.** Existing VTO models display limited flexibility in image editing, restricting their adaptability for practical applications. Essential functionalities—such as zero-shot generalization to previously unseen reference images, precise localized region editing and refinement, texture transfer, and object-level customization—are either only partially supported, absent, or require distinct models for each function (shown in Fig. 1b and Fig. 1d). These capabilities are essential in real-world VTO production environments where users expect tailored adjustments and high degrees of personalization, making them vital for enhancing the practical utility and user satisfaction in VTO applications. **5. A Universal Model for Diverse**

VTO Product Types. SOTA VTO models primarily focus on apparel, neglecting the diverse range of products customers may wish to virtually “try on,” including shoes, furniture, and accessories, etc (shown in Fig. 1b and Fig. 1f). This often necessitates the development of separate models for different product types, reducing generality and underscoring the need for a unified Virtual Try-All (VTA) solution that can cater to multiple product categories seamlessly.

In this work, we present DiT-VTON, a novel approach aimed at tackling above key challenges in the VTO domain. By leveraging the Diffusion Transformer (DiT)-based text-to-image (T2I) architecture, renowned for its high-fidelity, realistic image generation, we adapt this framework specifically to the image-conditioned VTO task. Our approach begins with a structured exploration of various model configurations to achieve optimal integration of reference and masked images. These configurations include: 1) In-Context token concatenation: treating the reference image and masked image (source image with overlaid mask) as additional tokens in the input token sequence. 2) Channel concatenation: concatenating the noised image, masked image, reference image and mask along the channel dimension. 3) Control network integration: using a separate control network to fuse conditions into the main network’s generation process. Additionally, we examine the enhancement of the generation process through pose conditions, achieved by stitching and concatenating pose information. Our evaluations reveal that token concatenation with pose stitching yields the best performance. To further boost robustness and versatility, we train the optimal configuration on an expanded dataset featuring in-the-wild examples, such as images with cluttered backgrounds, unstructured reference images, and product categories beyond garments. Experimental results indicate that scaling the dataset in this manner significantly enhances model’s performance on both wearable and non-wearable categories.

This paper makes four key contributions: 1) We examine various configurations of the Diffusion Transformer (DiT) to identify the optimal setup among them to take image conditions, enabling high-fidelity and efficient virtual try-on 2) We investigate different methods for incorporating pose control into the generation process, which further refines the model’s precision. 3) We demonstrate that scaling the training data with beyond-wearable categories, enhances VTO performance and enables the model to support virtual try-on across diverse product categories (virtual try-all). 4) Beyond basic try-on functionality, our model also supports image editing, including fine-grained adjustments and object-level customization, increasing its practical applications.

2. Related Work

Image Generation for Virtual Try-On Virtual try-on [9–11] refers to task that creates the image where a person nat-

urally wears the provided garment. Early efforts focus on Generative Adversarial Network (GAN) [12] with a two-stage generation. First, a warping model deforms the given garment into a shape that approximately fits the person’s pose. Then, a GAN-based generator fuses the wrapped clothing into the result image generation. The training focuses more on the pixel-level alignment and how to more effectively fuse the wrapping module into the main generator [13–16]. With the advent of more powerful Latent Diffusion Models [1–4], attention has shifted to how to effectively integrate the garment details into the person’s context, including dual U-Net [17], encouraging sparse attention over the clothing area [6], texture and human identity preservation [7, 18] and efficient training and inference [19]. In addition, VTO task is extended to multi-garment [20], multi-view [21], multi-modal [22], style controlled [23, 24], multi-category virtual try-all [9–11, 25, 26], mask-free try-on [27], etc.

Image Composition Virtual try-on can be considered as a special subset of image composition. Image composition uplifts the garment constraint in the reference image and the mask area in the source image to be clothing. It aims to detect the “transferable” part in the reference image and naturally blend it into the mask area. Built upon the text based image editing diffusion models, Paint by example [28] and Object stitch [29] approach this task by using a CLIP image encoder to extract the representation of the object. ControlNet-based model structure is further employed in Anydoor [30] to enhance the detail of the referenced objects. MimicBrush [31] uses Dual U-Nets and video samples to achieve non-object image customization.

Diffusion Transformer for Image Generation Diffusion Transformers replace the U-Net with transformer blocks and are proven to generate high-quality images [32]. FLUX models combine diffusion transformers with rectified flow achieve the state-of-the-art text to image generation and image editing abilities, meanwhile can significantly reduce the sampling steps. Though powerful, adapting Diffusion Transformers into inpainting tasks are rare. VITON-DiT [33] adapts diffusion transformers for video virtual try-on tasks. DiT [34] provides a concurrent work that trains FLUX for text-based inpainting tasks in a ControlNet style. M&M VTO [23] combines the U-Net and DiT for garment and person encoding separately.

3. Method

3.1. Background

3.1.1. Diffusion Transformer

Diffusion Transformers (DiTs) [32] introduce a new class of diffusion models that leverage transformer architecture for image generation. Diverging from traditional U-Net based approaches [1, 35], DiTs replaced U-Net module with vi-

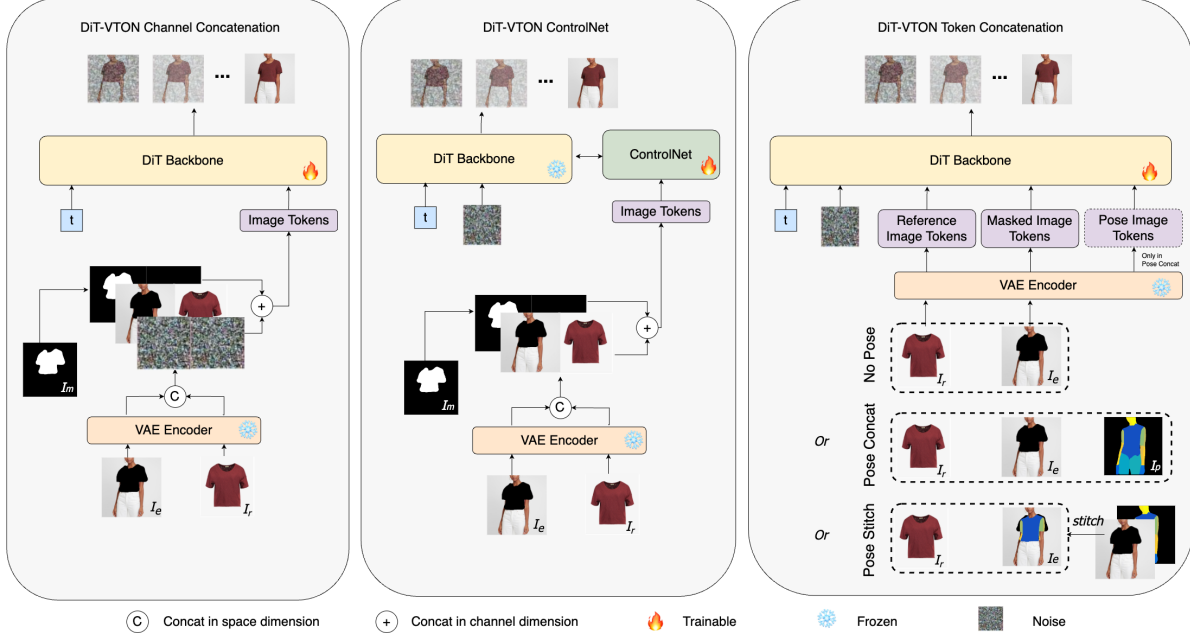


Figure 2. Illustration of different model configurations of DiT-VTON to effectively integrate image conditions.

sion transformer blocks. As demonstrated in [2], DiT architecture achieves both higher efficiency and higher quality sample generation.

3.1.2. Rectified Flow

Rectified Flow [36] aim to simplify the denoising process in Diffusion models by creating smooth, continuous transformations from noise to data in a more direct manner. By rectifying or adjusting the flow of noise toward data points, these models can potentially achieve faster and more efficient image synthesis. Mathematically, rectified flow models leverage differential equations to trace paths directly through high-dimensional data space. Let x_t represent the sample at time t within the continuous flow. The rectified field $\mathbf{f}_\theta(x_t, t)$ is trained to satisfy:

$$\frac{dx_t}{dt} = \mathbf{f}_\theta(x_t, t),$$

where $\mathbf{f}_\theta$ learns a trajectory that smoothly interpolates between Gaussian noise and the data distribution.

3.2. Diffusion transformer for Virtual Try-On and Virtual Try-All (DiT-VTON)

We formulate the VTO task as an image-conditioned inpainting problem. Given a source image I_s of a person wearing garment A, a reference image I_r of garment B, and a mask image I_m specifying the area to be edited resulting in $I_e = I_s \odot I_m$, the model’s objective is to generate a target image I_t where garment A in I_s is replaced with garment B from I_r . Each image input is first projected onto

the latent space $x = E(I)$ where $I \in \{I_s, I_m, I_r, I_e\}$, via an encoder E . Then, at each inference time step t , all the conditions along with the noise x_t are integrated together as a combined input to the denoising transformer backbone $x_{in} = f(x_t, x_r, x_m, x_e)$, where f represents different configurations to intake various image conditions by DiT.

3.2.1. Image Encoder

A typical choice in state-of-the-art VTO implementations is to use a dedicated garment encoder, which better preserves image features but also increases the number of model parameters, leading to higher computational costs and latency. As shown in the experimental results in subsequent sections, this added complexity is unnecessary; using a VAE encoder [37] directly can achieve comparable performance without any degradation in quality.

3.2.2. Image Condition Integration

We explore the optimal model configuration (f) to integrate additional image conditions into the transformer blocks. The architectures are shown in Fig. 2.

- **Channel Concatenation** (“Channel Concat”). We follow conventional approach in U-Net-based inpainting models [40] that concatenates the masked image I_e , the mask image I_m and the latent noise x_t in the channel dimension. As for the additional reference image, we concatenate it with the masked image at the spatial dimension, as similar practice in [7, 18, 19]: $x_c = x_e \oplus x_r$, $m_c = x_m \oplus \mathcal{O}$, and $z_c = x_t \oplus \mathcal{P}$, where $\oplus$ denotes spatial concatenation and $\mathcal{O}$ and $\mathcal{P}$ is the padding image. The

Table 1. Results of DiT-VTON’s different model configurations to integrate image conditions. (Sec. 3.2.2).

Model configuration	VITON-HD				DressCode			
	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	KID $\downarrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	KID $\downarrow$
DiT-VTON Channel Concat (<i>Ours</i>)	0.9065	0.0692	9.263	1.208	0.9250	0.0500	8.345	3.468
DiT-VTON ControlNet (<i>Ours</i>)	0.8930	0.0861	10.734	2.202	0.9242	0.0639	9.804	4.721
DiT-VTON Token Concat (<i>Ours</i>)	0.9130	0.0672	8.869	1.024	0.9379	0.0451	5.498	1.349

Table 2. Results of pose conditioning on best model configuration from Table 1, and comparison with SOTA VTO models.

Models	VITON-HD				DressCode			
	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	KID $\downarrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	KID $\downarrow$
StableVTON [6]	0.8543	0.0905	11.054	3.914	-	-	-	-
LaDI-VTON [38]	0.8603	0.0733	14.648	8.754	0.7656	0.2366	10.676	5.787
IDM-VTON [7]	0.8499	0.0603	9.842	1.123	0.8797	0.0563	9.546	4.320
OOTDiffusion [39]	0.8187	0.0876	12.408	4.680	0.8854	0.0533	12.567	6.627
CatVTON [19]	0.8704	0.0565	9.015	1.091	0.8922	0.0455	6.137	1.403
DiT-VTON Token Concat (<i>Ours</i>)	0.9130	0.0672	8.869	1.024	0.9379	0.0451	5.498	1.349
DiT-VTON Token Concat Pose Concat (<i>Ours</i>)	0.9229	0.0560	8.628	<u>0.881</u>	<u>0.9430</u>	<u>0.0397</u>	6.149	2.028
DiT-VTON Token Concat Pose Stitch (<i>Ours</i>)	<u>0.9216</u>	0.0576	<u>8.673</u>	0.820	0.9432	0.0389	<u>5.584</u>	1.540

final input to the transformer is $P(x_c \odot m_c \odot z_c)$, where $\odot$ is channel concatenation. To get the output, we only keep the first half of the output in the spatial dimension.

- **ControlNet.** Another conventional way of adding control to diffusion models is by copying (part of) the main denoising backbone as the ControlNet to encode conditions. Then the encoded image representation is fused back to the main backbone by cross attention, summation or other adaptive norm layers [32, 41]. Here, we concatenate x_c and m_c in the channel dimension as the input to ControlNet. z_c is the input to the main transformer blocks.
- **Token Concatenation** (“Token Concat”). Inspired by previous work [32, 42, 43], we patchify (denoted as P) each latent image into tokens and directly concatenate all the image tokens together as input: $P(x_t) \circ P(x_r) \circ P(x_e)$, where $\circ$ denotes concatenation in sequential dimension. After the final block, we remove the conditioning tokens from the sequence. Here, we drop the latent mask.

Adding pose constraint. In a concurrent work [9], we investigate methods to preserve pose information. In alignment with that work, here we explore two approaches to recover pose details hidden by the mask, by conditioning on a pose image I_p : 1) **Pose Concatenation** (“Pose Concat”): Adding the pose image tokens to the input sequence. While simple, it results in longer input with additional computation and inference latency. 2) **Pose Stitching** (“Pose Stitch”): Stitching the pose image into the masked area $I_e = I_s \odot I_m + I_p \odot (1 - I_m)$. Both methods are applied to the best performing approach, see Figure 2 right (Token Concatenation architecture) for an illustration of the two methods (bottom two rows of inputs). For more comprehensive experiments and analysis of pose conditioning

strategies, we refer the reader to [9].

4. Experiments

4.1. Datasets

Virtual Try-on dataset To evaluate the best model configuration for integrating image conditions in the VTO task, we conduct experiments using the test sets of two widely recognized public benchmarks: VITON-HD [13] and DressCode [44]. Each data pair in these datasets consists of a reference product image and a source image, where the source image serves as the target label, depicting the model wearing the product from the reference image. This setup allows the model to learn how to realistically overlay the product onto the source image. VITON-HD includes only upper garments in front-facing poses, with a total of 2,032 pairs for testing. For our experiments, we utilize the preprocessed mask and pose images provided with the dataset. DressCode offers a more diverse set of 5,400 test pairs, covering upper and lower garments as well as full-body dresses. For consistency in pose representation, we apply the same colormap as used in VITON-HD. To obtain clothing-agnostic masks, we use auto-masker prediction scripts. To alleviate the mask overfitting issue, we randomly sample 50% of the training data and use the minimal bounding box that covers the clothing-agnostic masks.

Virtual Try-all dataset Beyond the limitations of front-facing garment try-on and commonly used garment categories like shirts and pants, we collected a new dataset designed to cover more diverse clothing and product categories, as well as to handle more complex, in-the-wild scenarios. This dataset includes over 1,000 different object cat-

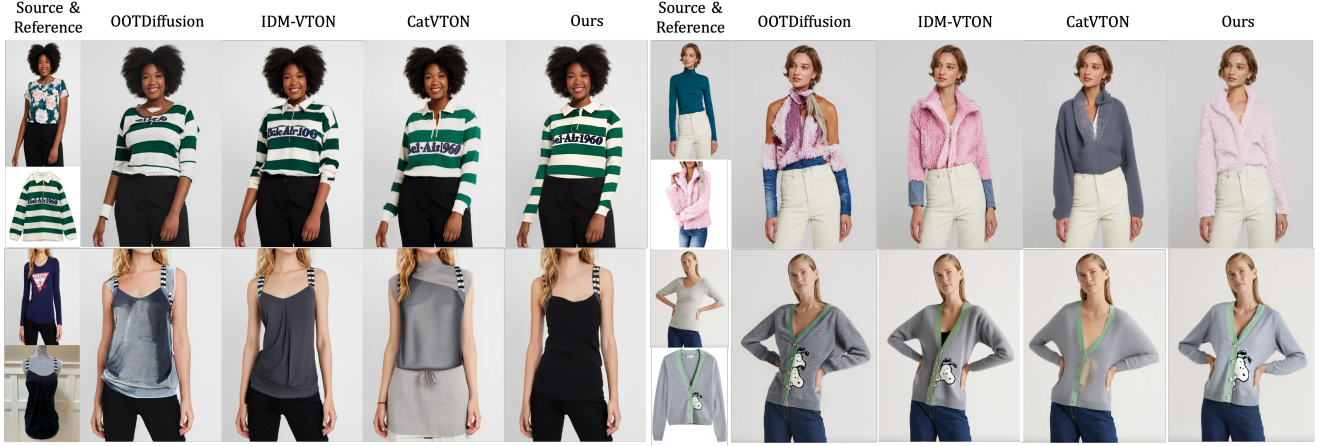


Figure 3. **Comparison of DiT-VTON with state-of-the-art models on virtual try-on test cases.** Our method can generate better garment details and layouts, e.g. number of buttons and puppy’s eyes/paws in bottom right example. Please zoom in to see details preservation.



Figure 4. **Comparison of DiT-VTON with state-of-the-art models on multi-garment try-on.** Our method can generate better garment details and layouts. Please zoom in to see details preservation.

egories and is structured as follows: 1) Expanded Clothing Categories: Examples include jackets, pajamas, and suits, etc; 2) Wearable and Non-Wearable Items: Wearable items include bags, hats, and shoes, while non-wearable items encompass jewelry, devices, toys, and more. 3) In-the-Wild Source and Reference Images: Source images may contain individuals in various background scenes, with subjects not necessarily facing forward. Reference images may not always present the product in isolation; for instance, clothing might appear in both front and back views or shown flat instead of modeled in 3D forms.

Data cleaning and preparation To ensure accurate image pair selection, we first apply a classifier to verify whether the source image is displaying the same product

shown in the reference image. For masks, we first apply a text-prompted object detection model to generate the bounding box area of the product in the source image using the product category as the text prompt. Then, we use the bounding box to prompt an object segmentation model to create the mask.

4.2. Implementation details

We use a diffusion transformer architecture as our diffusion backbone, building upon the implementation provided in the Diffusers library¹. All images are padded and resized to 512×512 pixels. The model is trained with a learning rate of $1e-5$ and a batch size of 32. During inference, we use a

¹<https://github.com/huggingface/diffusers>

sampling step of 28. The DiT model is tuned on 4 NVIDIA H100 80GB GPUs for 5,000 iterations.

4.3. Evaluation Metrics

The evaluation is conducted in both paired and unpaired settings. In the paired setting, the garment in the source image matches the garment in the reference image, with the source image serving as the ground truth. We use Structural Similarity Index (SSIM) [45] and Learned Perceptual Image Patch Similarity (LPIPS) [46] to compare the ground truth with the generated results. In the unpaired setting, the garment in the source image differs from that in the reference image. Here, we compute the distributional distance between the source and generated images using Fréchet Inception Distance (FID) [46] and Kernel Inception Distance (KID) [47]. We follow the implementation in [38].

4.4. Quantitative Results

In Sec. 4.4.1 - Sec. 4.4.3, we quantitatively demonstrate that token concatenation with pose stitching (both defined in Sec. 3.2.2) yield our best-performing model configuration. Additionally, we evaluate its generalization across diverse product categories in Sec. 4.4.4.

4.4.1. Comparison between model configurations

We evaluate all 3 proposed conditional model configurations (Sec. 3.2.2) to identify the optimal image integration setup for VTO task among the explored ones. Tab. 1 presents the evaluation results on the VITON-HD and DressCode datasets. Our findings indicate that using concatenated image tokens as model inputs yields the best performance. This approach is closely followed by channel concatenation, which outperforms the ControlNet-based structure. Notably, this performance hierarchy remains consistent across both the VITON-HD and DressCode datasets.

4.4.2. Comparison with baseline SOTA models

We compare our approach with state-of-the-art methods as outlined in [19], a comprehensive and recent study in the field. As shown in Tab. 2, our model, which employs image token concatenation, significantly outperforms the established VTO models across all evaluated metrics on both the VITON-HD and DressCode datasets.

4.4.3. Pose preservation

In addition, in Tab. 2, we demonstrate the impact of adding pose conditioning to our models and compare two integration strategies. Overall, incorporating pose condition enhances model performance, although it introduces a slight decrease in unpaired generalization (as indicated by the KID metric) on the DressCode dataset. When comparing the two integration strategies—pose concatenation and pose stitching—we observe that their performances are closely comparable. Specifically, pose concatenation yields better

results on the VITON-HD dataset, whereas pose stitching is more effective on the DressCode dataset.

Table 3. **Quantitative results of DiT-VTON compared with SOTA models on general product categories (Virtual Try-All).** AnyDoor/Mimicbrush[original] denotes open-source publicly available checkpoints, while AnyDoor/Mimicbrush[vital] represents training the respective model on our Virtual Try-All dataset (as an ablation on both the model/method and the dataset).

Model [Training Dataset]	Non-wearable		Wearable	
	SSIM $\uparrow$	LPIPS $\downarrow$	SSIM $\uparrow$	LPIPS $\downarrow$
AnyDoor[original] [30]	0.7313	0.2954	0.7442	0.2787
AnyDoor[vital]	0.7766	0.2324	0.7857	0.2157
MimicBrush[original] [31]	0.9060	0.0924	0.8852	0.1100
MimicBrush[vital]	0.9078	0.0910	0.8998	0.0896
DiT-VTON[vton] (<i>Ours</i>)	<u>0.9088</u>	<u>0.0898</u>	<u>0.9082</u>	<u>0.0731</u>
DiT-VTON[vital] (<i>Ours</i>)	0.9281	0.0617	0.9171	0.0617

4.4.4. Generalization beyond garment try-on

Beyond basic garment try-on, we evaluate the generalization capabilities of our best model configuration, token concatenation with pose stitching, using our in-house collected Virtual Try-All dataset. First, we assess the model in a zero-shot scenario, training it exclusively on garment categories (denoted as DiT-VTON[vton]) and comparing its performance against Anydoor [30] and MimicBrush [31] on both wearable and non-wearable test sets. As shown in the first section of Tab. 3, DiT-VTON[vton] demonstrates notable zero-shot capabilities, even when trained solely on garment categories. Next, we extend the training set to include a broader range of product categories and re-evaluate the model on the Virtual Try-All dataset (refer to the last row in Tab. 3). This expanded model, denoted DiT-VTON[vital] outperforms the VTON-specific model across both test sets. Additionally, we present results obtained via training AnyDoor/MimicBrush with our Virtual Try-All dataset (denoted as AnyDoor/MimicBrush[vital]) as an ablation of the dataset from the model. These results show that scaling and diversifying the training data helps with model performance across categories, both wearable and non-wearable.

4.5. Qualitative results

In this section, we present qualitative examples to illustrate the superior robustness, generalization capabilities, versatility, and strong image customization abilities of our DiT-VTON model. All results are generated using our best-performing model configuration (namely DiT-VTON Token Concat Pose Stitch), as quantitatively evidenced in Sec. 4.4.

4.5.1. Comparison with baseline models

Using in-the-wild test examples, we provide pairwise comparisons between the outputs of our method and those of



Figure 5. **Comparison of DiT-VTON with state-of-the-art models on generalization ability of virtual try-all (VTA)/image editing.** Our method can generate better garment details and layouts, e.g. the texture of vase in bottom left example. Please zoom in to see details.



Figure 6. **Comparison of DiT-VTON model configuration with and without pose conditioning.** See more figures of pose preservation experiments in the Supplementary Material.

baseline models in Fig. 3. As shown, our model consistently excels in preserving key attributes: texture in the striped shirt, styling in the pink coat, color in the black sleeveless shirt, and details in the cartoon character graphic. We further evaluate our model in a multi-garment try-on scenario. Using a full-body outfit reference image, our model accurately captures the clothing “semantics” from the reference and generates high-fidelity results for lower-body, upper-body, and full-body try-on.

4.5.2. Generalization beyond garment try-on

Our model’s performance also generalizes to categories beyond garments, such as hats, handbags, jewelry, furniture, etc. As illustrated in Fig. 5, compared to Anydoor [30] and

MimicBrush [31], our model generates more photorealistic and style-preserved results on bags and shoes. Furthermore, as shown in the second row in Fig. 5, our model outperforms the baseline models on local editing and texture transfer.

4.5.3. Pose preservation

In Fig. 6, we compare results generated by our DiT-VTON model from the same source image, with and without pose conditioning. This comparison illustrates that pose conditioning avoids hallucination of arm poses and hands and helps with the preservation of source pose, consistent with the quantitative findings.

5. Conclusion and Limitations

In this paper, we explored the use of Diffusion Transformer for the Virtual Try-on (VTO) task. We investigated three model configurations—Token Concatenation, Channel Concatenation, and ControlNet—for integrating image conditions, along with two pose preservation strategies: Pose Concatenation and Pose Stitching. Our experimental results demonstrate that Token Concatenation with Pose Stitching consistently outperforms state-of-the-art methods on two widely used public datasets, VITON-HD and Dress-Code. Additionally, we propose an innovative redefinition of VTO, broadening it beyond basic “garment try-on” to a more versatile and personalized task. Building on this concept, we present DiT-VTON, an all-purpose virtual try-on model that not only generalizes to diverse product categories for “Virtual Try-all” (VTA) applications, but is also integrated with multiple image customization abilities including pose preservation, precise localized region editing and refinement, texture transfer and object-level customization. Our model quantitatively and qualitatively outper-

forms state-of-the-art VTO models across all functionalities and demonstrates enhanced robustness when applied to in-the-wild and even low-resolution images. As a limitation, DiT-VTON may struggle with certain product categories, particularly those that occupy a small area within reference images. For these cases, the generated results tend to show blurry artifacts around the mask edges.

References

- [1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 3
- [2] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 4
- [3] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.
- [4] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 1, 3
- [5] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2
- [6] Jeongho Kim, Guojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. 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*, pages 8176–8185, 2024. 2, 3, 5
- [7] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for virtual try-on. *arXiv preprint arXiv:2403.05139*, 2024. 2, 3, 4, 5
- [8] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2
- [9] Qi Li, Shuwen Qiu, Julien Han, Xingzi Xu, Mehmet Saygin Seyfioglu, Kee Kiat Koo, and Karim Bouyarmane. Efficient encoder-free pose conditioning and pose control for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. AI for Content Creation Workshop. 3, 5
- [10] Xingzi Xu, Qi Li, Shuwen Qiu, Julien Han, and Karim Bouyarmane. Deft-vton: Efficient virtual try-on with consistent generalised h-transform. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3308–3317, 2025.
- [11] Julien Han, Shuwen Qiu, Qi Li, Xingzi Xu, Mehmet Saygin Seyfioglu, Kavosh Asadi, and Karim Bouyarmane. Instructvton: Optimal auto-masking and natural-language-guided interactive style control for inpainting-based virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. AI for Content Creation Workshop. 3
- [12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3
- [13] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14131–14140, 2021. 3, 5
- [14] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8485–8493, 2021.
- [15] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *European Conference on Computer Vision*, pages 204–219. Springer, 2022.
- [16] Zhenyu Xie, Zaiyu Huang, Xin Dong, Fuwei Zhao, Haoye Dong, Xijin Zhang, Feida Zhu, and Xiaodan Liang. Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23550–23559, 2023. 3
- [17] Luyang Zhu, Dawei Yang, Tyler Zhu, Fitsum Reda, William Chan, Chitwan Saharia, Mohammad Norouzi, and Ira Kemelmacher-Shlizerman. Tryondiffusion: A tale of two unets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4606–4615, 2023. 3
- [18] Xu Yang, Changxing Ding, Zhibin Hong, Junhao Huang, Jin Tao, and Xiangmin Xu. Texture-preserving diffusion models for high-fidelity virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7017–7026, 2024. 3, 4
- [19] Zheng Chong, Xiao Dong, Haoxiang Li, Shiyue Zhang, Wenqing Zhang, Xujie Zhang, Hanqing Zhao, and Xiaodan Liang. Catvton: Concatenation is all you need for virtual try-on with diffusion models. *arXiv preprint arXiv:2407.15886*, 2024. 3, 4, 5, 7
- [20] Yuhan Li, Hao Zhou, Wenxiang Shang, Ran Lin, Xuanhong Chen, and Bingbing Ni. Anyfit: Controllable virtual try-on for any combination of attire across any scenario. *arXiv preprint arXiv:2405.18172*, 2024. 3
- [21] Haoyu Wang, Zhilu Zhang, Donglin Di, Shiliang Zhang, and Wangmeng Zuo. Mv-vton: Multi-view virtual try-on with diffusion models. *arXiv preprint arXiv:2404.17364*, 2024. 3

- [22] Xujie Zhang, Ente Lin, Xiu Li, Yuxuan Luo, Michael Kampffmeyer, Xin Dong, and Xiaodan Liang. Mmtryon: Multi-modal multi-reference control for high-quality fashion generation. *arXiv preprint arXiv:2405.00448*, 2024. 3
- [23] Luyang Zhu, Yingwei Li, Nan Liu, Hao Peng, Dawei Yang, and Ira Kemelmacher-Shlizerman. M&m vto: Multi-garment virtual try-on and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1346–1356, 2024. 3
- [24] Mengting Chen, Xi Chen, Zhonghua Zhai, Chen Ju, Xuewen Hong, Jinsong Lan, and Shuai Xiao. Wear-any-way: Manipulable virtual try-on via sparse correspondence alignment. *arXiv preprint arXiv:2403.12965*, 2024. 3
- [25] Mehmet Saygin Seyfioglu, Karim Bouyarmane, Suren Kumar, Amir Tavaneai, and Ismail B Tutar. Diffuse to choose: Enriching image conditioned inpainting in latent diffusion models for virtual try-all. *arXiv preprint arXiv:2401.13795*, 2024. 3
- [26] Mehmet Saygin Seyfioglu, Karim Bouyarmane, Suren Kumar, Amir Tavaneai, and Ismail B Tutar. Dreampaint: Few-shot inpainting of e-commerce items for virtual try-on without 3d modeling. *arXiv preprint arXiv:2305.01257*, 2023. 3
- [27] Xuanpu Zhang, Dan Song, Pengxin Zhan, Qingguo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Anan Liu. Boov-ton: Boosting in-the-wild virtual try-on via mask-free pseudo data training. *arXiv preprint arXiv:2408.06047*, 2024. 3
- [28] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023. 3
- [29] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Object-stitch: Object compositing with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18310–18319, 2023. 3
- [30] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6593–6602, 2024. 3, 7, 8
- [31] Xi Chen, Yutong Feng, Mengting Chen, Yiyang Wang, Shilong Zhang, Yu Liu, Yujun Shen, and Hengshuang Zhao. Zero-shot image editing with reference imitation. *arXiv preprint arXiv:2406.07547*, 2024. 3, 7, 8
- [32] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 3, 5
- [33] Jun Zheng, Fuwei Zhao, Youjiang Xu, Xin Dong, and Xiaodan Liang. Viton-dit: Learning in-the-wild video try-on from human dance videos via diffusion transformers. *arXiv preprint*, 2024. 3
- [34] <https://github.com/alimama-creative/flux-controlnet-inpainting>. 2024. 3
- [35] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 3
- [36] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 4
- [37] Lucas Pinheiro Cinelli, Matheus Araújo Marins, Eduardo Ant3nio Barros da Silva, and S3rgio Lima Netto. Variational autoencoder. In *Variational Methods for Machine Learning with Applications to Deep Networks*, pages 111–149. Springer, 2021. 4
- [38] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8580–8589, 2023. 5, 7
- [39] Yuhao Xu, Tao Gu, Weifeng Chen, and Chengcai Chen. Oot-diffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. *arXiv preprint arXiv:2403.01779*, 2024. 5
- [40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022. 4
- [41] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 5
- [42] <https://github.com/black-forest-labs/flux>. 2024. 5
- [43] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024. 5
- [44] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: High-resolution multi-category virtual try-on. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2231–2235, 2022. 5
- [45] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7
- [46] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7
- [47] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 7

DiT-VTON: Diffusion Transformer Framework for Unified Multi-Category Virtual Try-On and Virtual Try-All with Integrated Image Editing

Supplementary Material

6. More Qualitative Examples

We provide qualitative examples of in-the-wild try-on, multiple garments try-on, try-on texture transfer and try-on local editing to compare our models with baselines including OOTDiffusion, IDM-VTON and CatVTON in Figs. 7 to 14. We provide qualitative examples of virtual try all to compare our models with AnyDoor and MimicBrush in Figs. 15 to 18. In Fig. 19, we provide additional virtual try on examples with pose control.

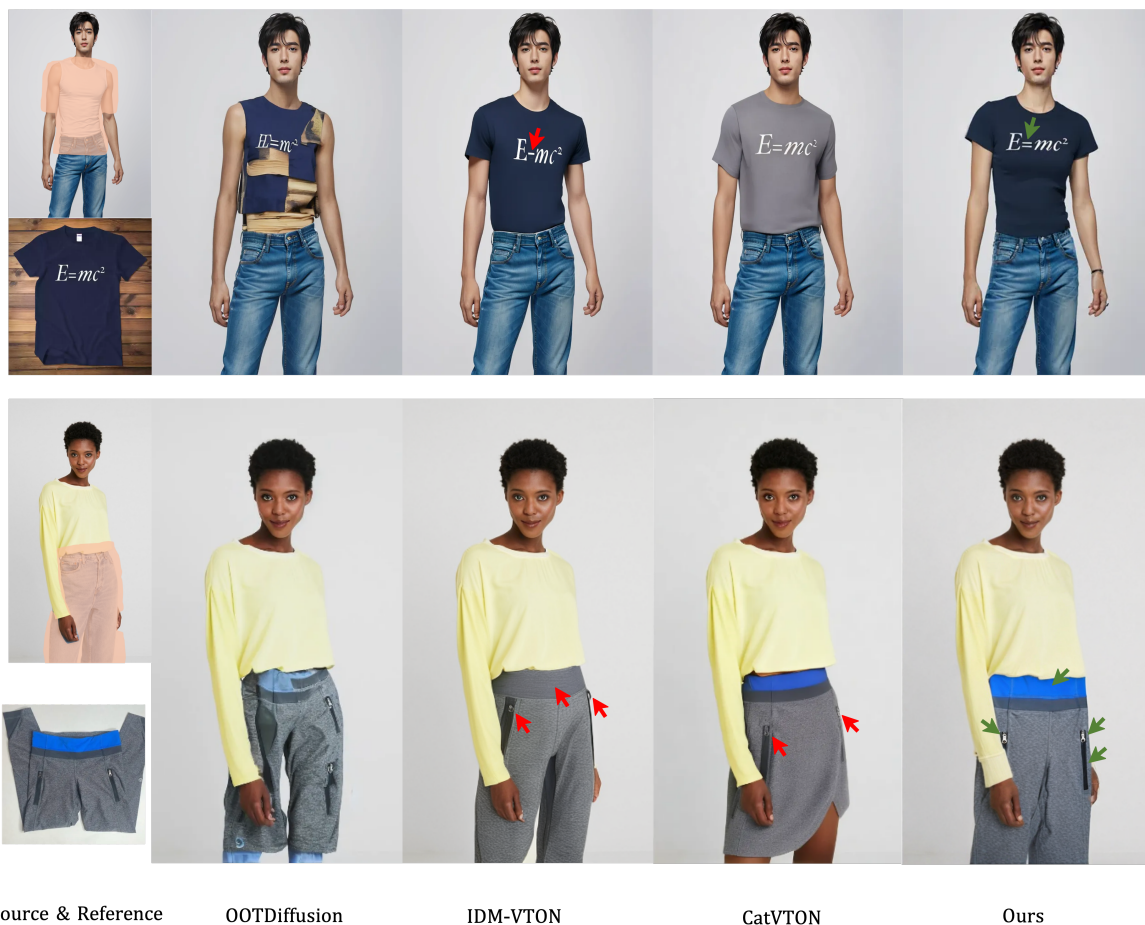


Figure 7. **In-the-wild try on: unorganized reference image and text preservation.** Red arrow highlights error of alternative state-of-the-art models. Green arrow highlights correct output from our model. Our method can generate better garment details and layouts. Please zoom in to see details preservation.

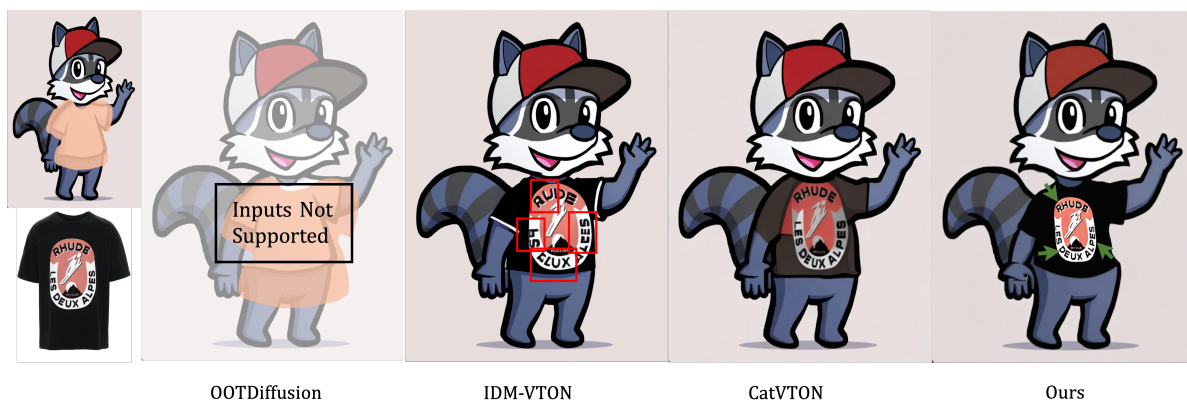


Figure 8. **In-the-wild try on: in-the-wild source image and text preservation.** Red boxes highlight errors of alternative state-of-the-art models. Green arrow highlights correct output from our model. Our method can better preserve subtle garment details and layouts. Please zoom in to see details preservation.



Source & Reference

OOTDiffusion

IDM-VTON

CatVTON

Ours

Figure 9. **Multi-garment try on.** Red arrow highlights less clarity details generated from alternative state-of-the-art models. Green arrow highlights correct output from our model. Our method can generate better and higher clarity garment details and layouts. Please zoom in to see details.

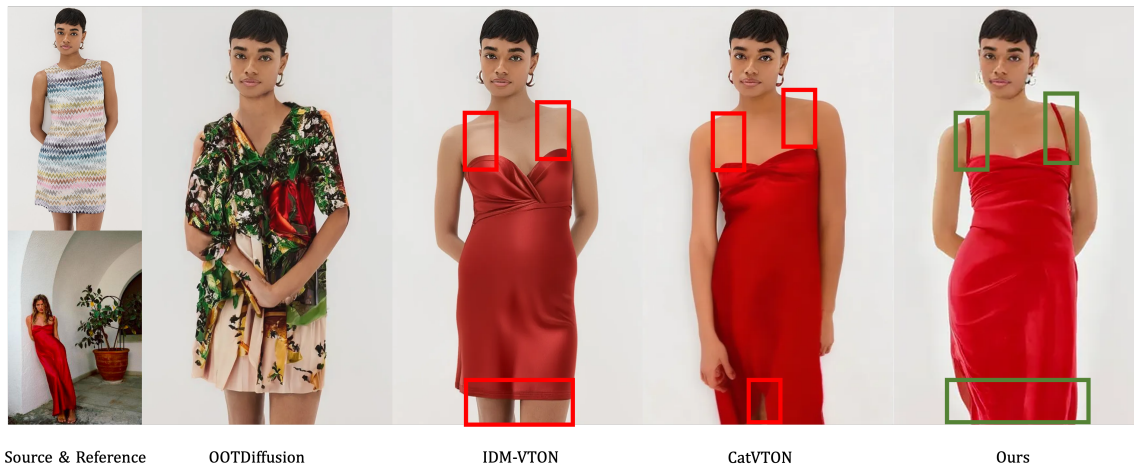


Figure 11. **In-the-wild try on: person-wearing reference image.** Red boxes highlight errors of alternative state-of-the-art models. Green boxes highlight correct output from our model.

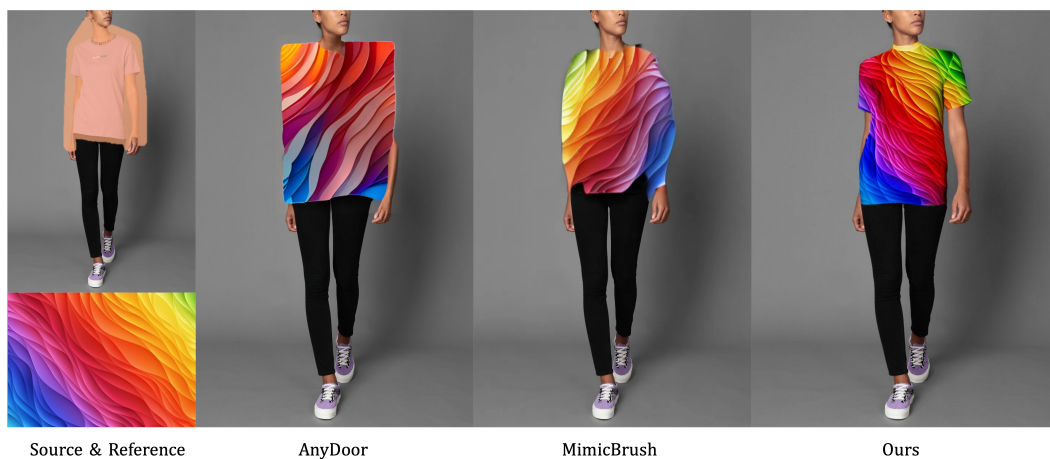


Figure 12. **Image customization: texture transfer.**



Figure 13. **Image customization: local editing.**



Figure 14. **Additional Virtual try on (VTO) customization examples.** Red boxes highlight errors of alternative state-of-the-art models. Green boxes highlight correct output from our model. Our method can generate better garment details and layouts. Please zoom in to see details preservation.



Figure 15. **Virtual try all (VTA) - Wearable: necklace.**



Figure 16. **Virtual try all (VTA) - Wearable: scarf.**



Figure 17. **Virtual try all (VTA) - Wearable: hat.**



Reference

Source

AnyDoor

MimicBrush

Ours

Figure 18. **Virtual try all (VTA) - Non-Wearable.** Our method can generate better product details and layouts. Please zoom in to see details preservation.

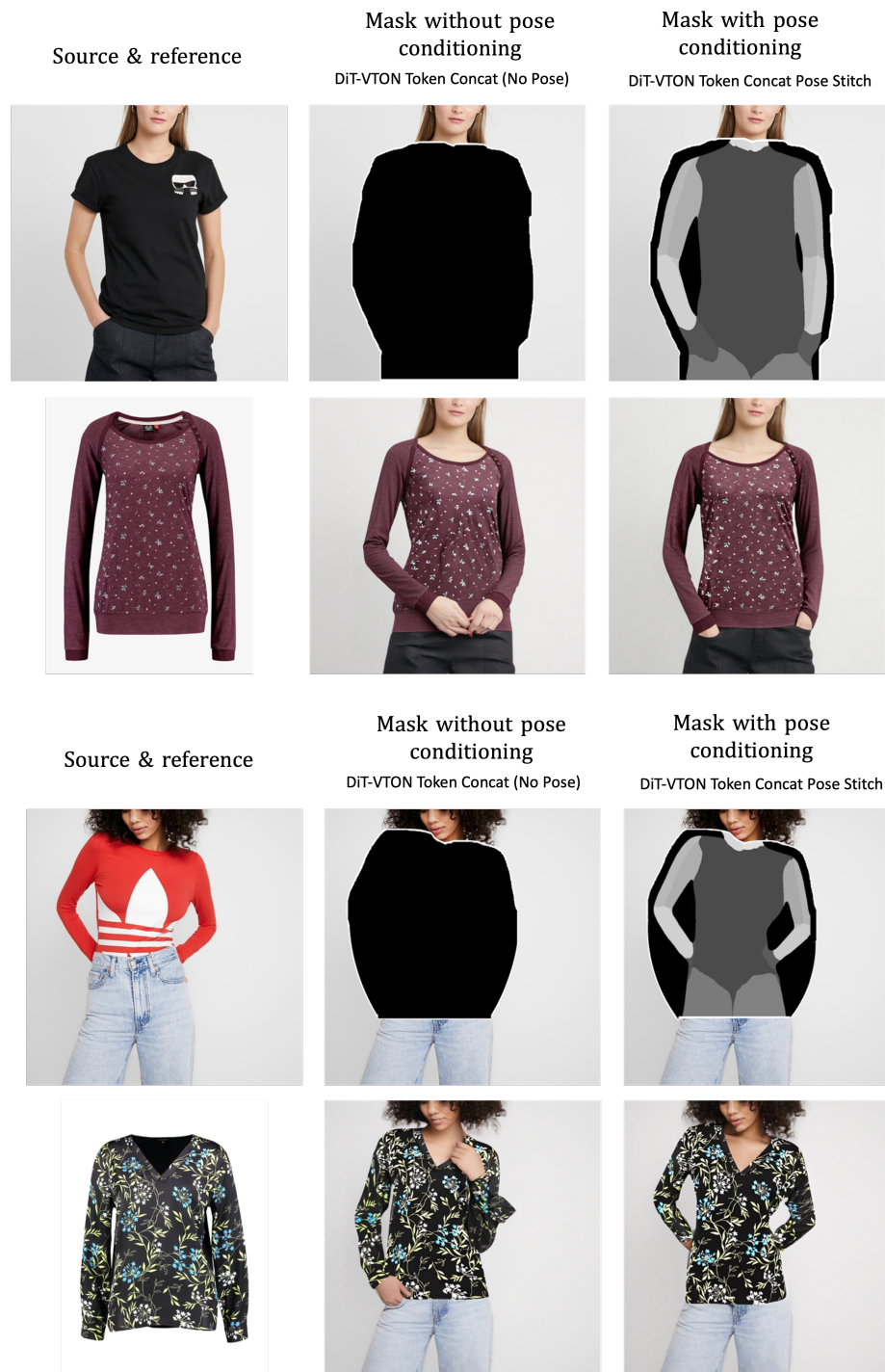


Figure 19. In-the-wild try on: pose preservation.