

Efficient Encoder-Free Pose Conditioning and Pose Control for Virtual Try-On

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://pose-vton.github.io/vto-pose-conditioning/>

Work submitted in November 2024 to CVPR 2025

Abstract

As online shopping continues to grow, the demand for Virtual Try-On (VTON) technology has surged, allowing customers to visualize products on themselves by overlaying product images onto their own photos. An essential yet challenging condition for effective VTON is pose control, which ensures accurate alignment of products with the user’s body while supporting diverse orientations for a more immersive experience. However, incorporating pose conditions into VTON models presents several challenges, including selecting the optimal pose representation, integrating poses without additional parameters, and balancing pose preservation with flexible pose control.

In this work, we build upon a baseline VTON model that concatenates the reference image condition without external encoder, control network, or complex attention layers. We investigate methods to incorporate pose control into this pure-concatenation paradigm by spatially concatenating pose data, comparing performance using pose maps and skeletons, without adding any additional parameters or module to the baseline model. Our experiments reveal that pose stitching with pose maps yields the best results, enhancing both pose preservation and output realism. Additionally, we introduce a mixed-mask training strategy using fine-grained and bounding box masks, allowing the model to support flexible product integration across varied poses and conditions.

Our contributions are threefold: 1) We explore different configurations for integrating pose representations into VTON models, 2) We propose a lightweight, parameter-efficient approach for adding pose control, and 3) We en-

able flexible pose generation through mixed-mask training. Evaluations on public benchmarks demonstrate that our method improves pose preservation and outperforms state-of-the-art models with more complex conditioning frameworks, advancing VTON’s adaptability and realism for diverse real-world applications.

1. Introduction

As online shopping grows in popularity, so does the demand for customers to “try on” products virtually by superimposing them onto their own photos. This process, known as Virtual Try-On (VTON) [1–3], involves using a source image provided by the user, a specified mask area (of the source image), and an image of the desired product to try on (reference image). The model then seamlessly integrates the product into the masked area of the user provided source image, as shown in Fig. 1. Recent advances in UNet-based Latent Diffusion Models (LDMs) [4–7] have enabled VTON to be formulated as an image-conditioned inpainting problem, which is fine-tuned with diffusion-based pre-trained weights and has proven to achieve promising generation quality.

A crucial aspect of VTON tasks is pose control. Pose information is essential for VTON for several reasons. First, when the mask obscures parts of the user’s pose in the source image, the model may “hallucinate” limbs or other body parts, which is particularly challenging for generative models to render realistically—especially with hands. This often results in less satisfactory or realistic outputs, as seen in Fig. 2. Additionally, incorporating pose control enables VTON models to produce results with various user poses and orientations, enhancing the diversity and realism of

*Equal contribution.

†Work done during internship at Amazon.



Figure 1. **Pose preservation and control.** In the first row, pose preservation requires the model to generate natural try-on results while maintaining the user’s original pose, even when parts of the body are obscured by a mask. In the second row, pose control tasks the model with generating images of the user in different poses, guided by an input pose image. All examples shown are generated by our model.

user experiences. Although crucial, integrating poses into VTON models presents several challenges: 1) **Pose Representation:** Common pose representations include pose maps (e.g. [8]) and skeletons (e.g. [9]). Current VTON approaches may use either representation or combine them, but it remains unclear which works best for VTON tasks. 2) **Integration with Model Inputs:** Current models add pose conditions as additional input channels, which requires modifications to the original model weights and additional training parameters. This raises the question of whether pose conditions can be incorporated without extra parameters. 3) **Pose Control vs. Pose Preservation:** In addition to maintaining the user’s original pose, VTON models must also support customization by generating try-on results with varying user poses. Unlike general image editing, VTON models generate within a masked area, which can hinder realistic outputs if the mask is not aligned with the pose.

In this work, we address these challenges by building

on efficient VTON models like DiT-VTON and DEFT-VTON [1–3], a paradigm that simplified to the extreme the VTON task by removing any need for external image encoders, additional encoding networks (“garment networks”, “reference networks”, “person network” etc), or complex cross-attention layers. DiT-VTON simply concatenates the masked source image and the reference image along the spatial dimension and feeds the concatenated result into the main diffusion network. This approach achieves simpler and more efficient training while maintaining high-quality generation and detail preservation, outperforming complex architecture alternatives. Motivated by the performance, we investigate whether additional conditions like poses can be further concatenated in the spatial dimension without introducing extra parameters. We use the pose conditioning signal as the additional signal, due to the observed limitation of baseline model in this domain (pose hallucination, hand hallucination when hand is not visible in the reference,



Figure 2. A hard case for pose preservation: When body parts fall within the masked area, the model must hallucinate the pose, and cause natural hand appearance.

etc). We evaluate both pose concatenation and pose stitching techniques, as well as their performance with pose maps and skeletons. Our results indicate that pose stitching with pose maps produces the best outcomes.

Furthermore, we train our model using both fine-grained masks and bounding box masks, enabling it to adapt flexibly to diverse mask conditions. The model learns to fit products precisely within fine-grained masks while allowing more versatile generation with bounding box masks, supporting varied pose-controlled outputs.

Our contributions unfold in three dimensions: 1) We systematically study configurations for integrating different pose representations into VTON models; 2) We propose a simple yet efficient method to incorporate pose conditions without additional channels or extra model parameters; and 3) We enable flexible pose generation through a mixed-mask training strategy. Evaluations on public test benchmarks VITON-HD [10] and DressCode [11] demonstrate that our method enhances pose preservation and surpasses state-of-the-art (SOTA) VTON models that rely on more complex pose-conditioning frameworks.

2. Related Work

Image Virtual Try-On with Diffusion Models Latent Diffusion Models (LDMs) [4–7] have demonstrated strong generative performance across tasks such as text-to-image generation, image inpainting, and image editing. To adapt diffusion models for VTON tasks, WarpDiffusion [12] integrates a warping module [13–16] that aligns the garment with the masked area and uses a pre-trained text-to-image diffusion model enhanced by a novel information and local garment feature attention mechanism. TryOn-

Diffusion [17] introduces a dual-UNet structure with cross-attention to implicitly incorporate the warping process, linking the garment condition with the model. For better garment representation infusion into the denoising UNet, LaDI-VTON [18] applies a textual inversion module to embed garment-specific information.

Moving beyond warping, StableVTON [19] incorporates a supervision signal that aligns the attention map with the warped clothing, while IDM-VTON [20] combines textual inversion and a dual-UNet structure to achieve robust try-on results with diverse, real-world images. Further extensions in VTON, such as Anyfit [21], MV-VTON [22], and Wear-Any-Way [23], introduce capabilities for multi-garment, multi-view, and interactive editing. Meanwhile, some works shift towards a single UNet structure. CatVTon [24] reduces complexity by removing extra garment modules and cross-attention layers; TPD [25] uses a single UNet with an additional mask channel to refine the mask input; MMTryon [26] leverages additional encoders for multi-modal, multi-reference generation, while M&M VTO [27] incorporates a Diffusion Transformer and supplementary encoders to enable style-controlled generation.

Adding Control to Diffusion Models Diffusion models for text based image editing have shown superior performances. However, abstract text prompting cannot provide details about visual concepts such as poses, gestures, textures, etc. A conventional way to introduce control signals into the diffusion models involves using a copy of the UNet encoder as the ControlNet [28], which encodes the control images such as sketches, poses, and depth information into the denosing UNet. Text inversion with IP Adapter [29] and DreamBooth [30] offers another alternative for additional conditioning. In VTON task specifically, pose signal is an important factor to help preserve and generate various poses for the given subject. Most works add additional channels to accommodate the pose images [18–20, 22, 25, 31]. Meanwhile, two types of pose representations — pose maps (e.g. [8]) and skeletons (e.g. [9]) — are mixedly used without validations. In this work, we take a closer look at the pose conditioning and study the performance of different pose representations and the effective and efficient way to incorporate pose signals into the generation process.

3. Method

3.1. Preliminary

3.1.1. Latent Diffusion Models

Latent Diffusion Models [4, 32] operate by learning a gradual denoising process through a series of timesteps that reverse a noise-injection process. Given an image x_0 , the forward diffusion process adds Gaussian noise step-by-step, creating a latent variable x_t at each timestep t . This process

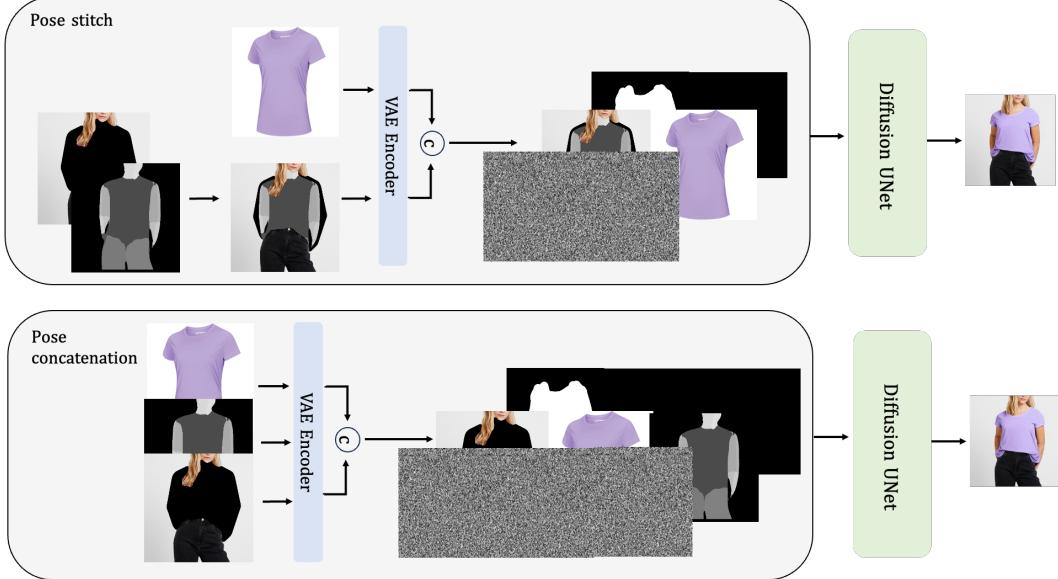


Figure 3. **Adding pose conditions.** The top image illustrates the construction of the input using pose stitching, while the bottom image shows the model input with pose concatenation.

is represented as:

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t} \epsilon,$$

where α_t is a noise schedule parameter and $\epsilon \sim \mathcal{N}(0, I)$ is Gaussian noise. The model is then trained to predict the noise ϵ and gradually denoise x_t back to x_0 over multiple steps. Stable Diffusion optimizes this process by working within a compressed latent space, enabling high-resolution generation with reduced computational requirements.

3.1.2. Denoising Diffusion Implicit Models (DDIM)

DDIM [33] builds on diffusion models by introducing a deterministic sampling method that accelerates the reverse diffusion process. Unlike traditional diffusion, which follows a Markovian sequence, DDIM allows for non-Markovian sampling with fewer steps. The reverse process in DDIM can be written as:

$$x_{t-1} = \sqrt{\alpha_{t-1}} \hat{x}_0 + \sqrt{1 - \alpha_{t-1}} \epsilon_t,$$

where \hat{x}_0 is an estimate of the original image at each timestep. This approach enables DDIM to achieve high-quality results with significantly fewer sampling steps, enhancing efficiency while maintaining image fidelity.

3.2. Adding Pose Conditions into the Diffusion Models

Virtual try-on generates a composite image of a person wearing a specified garment. The input includes a person image $I_p \in \mathbb{R}^{3 \times H \times W}$, a garment image $I_g \in \mathbb{R}^{3 \times H \times W}$,

and a binary mask M defining the editable area. The inpainted person image, which serves as input, is denoted as $I_m = I_p \otimes M$. To enable precise pose control, a pose image I_c is additionally extracted from I_p . These inputs— I_m , I_g , and I_c —are encoded into the latent space by a VAE encoder E : $x_m = E(I_m)$, $x_g = E(I_g)$, $x_c = E(I_c)$ where $x \in \mathbb{R}^{4 \times H/8 \times W/8}$. The mask M is interpolated to the size $m \in \mathbb{R}^{H/8 \times W/8}$ to align with the latent representations.

3.2.1. Stitching Conditions in the Spatial Dimension

As shown in recent work [1, 20, 24, 25, 34], stitching a reference image with the masked image along the spatial dimension can effectively preserve text and logo details in garments. Specifically, Stable Diffusion (SD) inpainting models extend the original 4-channel noisy image into 9 channels by concatenating z_t , x_m , and m in the channel dimension, denoted as concat_c . To integrate an additional garment image x_g , it is concatenated alongside the masked image in the spatial dimension, denoted as concat_s . Additional channels are padded with black images $x_\alpha \in \mathbb{R}^{4 \times H/8 \times W/8}$ and $x_\beta \in \mathbb{R}^{H/8 \times W/8}$: $x_{\text{masked}} = \text{concat}_s([x_m, x_g])$, $x_{\text{mask}} = \text{concat}_s([m, x_\beta])$, $x_{\text{noisy}} = \text{concat}_s([x_t, x_\alpha])$.

The final input to the transformer blocks is $x_{\text{in}} = \text{concat}_c[x_{\text{noisy}}, x_{\text{masked}}, x_{\text{mask}}]$. The model’s output is then cut in half.

3.3. Adding Pose Constraints

Above, we considered the minimal conditions for the VTON task. Additional conditions, such as pose and depth information, can help better preserve the pose within the

Table 1. Quantitative comparison on VTION-HD dataset of different methods for combining pose representation (joints vs. pose maps), integration methods (stitching vs. concatenation), and color modes (color vs. grayscale).

Pose conditions	Stable Diffusion v1.5				Stable Diffusion XL			
	SSIM↑	LPIPS↓	FID↓	KID↓	SSIM↑	LPIPS↓	FID↓	KID↓
Ours w/ Joints Stitch	0.8861	<u>0.0958</u>	<u>9.052</u>	<u>1.175</u>	<u>0.8864</u>	0.0941	<u>10.091</u>	<u>2.088</u>
Ours w/ Joints Concat	0.8781	0.1087	10.467	1.725	0.8695	0.1078	12.517	3.734
Ours w/ Pose Concat	0.8870	0.1148	12.623	3.277	0.8861	0.0928	11.895	3.144
Ours w/ Pose Concat Gray	<u>0.8879</u>	0.1120	12.449	3.256	0.8856	<u>0.0918</u>	11.809	2.975
Ours w/ Pose Stitch Gray	0.9053	0.0694	8.646	0.872	0.8993	0.0766	9.341	1.367

Table 2. Quantitative comparison with baselines on the VITON-HD and DressCode test sets. Models labeled with /w SD and SDXL use Stable Diffusion v1.5 and Stable Diffusion XL as the backbone, respectively.

Models	VITONHD				DressCode			
	SSIM↑	LPIPS ↓	FID↓	KID↓	SSIM↑	LPIPS ↓	FID↓	KID↓
StableVTON [19]	0.8543	0.0905	11.054	3.914	-	-	-	-
LaDI-VTON [18]	0.8603	0.0733	14.648	8.754	0.7656	0.2366	10.676	5.787
IDM-VTON [20]	0.8499	<u>0.0603</u>	9.842	1.123	0.8797	0.0563	9.546	4.320
CatVTON [24]	0.8704	0.0565	<u>9.015</u>	<u>1.091</u>	0.8922	0.0455	6.137	1.403
Ours w/ SD	0.9053	0.0694	8.646	0.872	0.9277	<u>0.0510</u>	5.103	0.951
Ours w/ SDXL	<u>0.8993</u>	0.0766	9.341	1.367	0.9252	0.0532	<u>5.545</u>	<u>1.358</u>

masked area. In practice, there are two primary representations of pose: pose maps and skeletons. A pose map is a dense representation, often visualized as a heatmap or color-coded overlay, where each pixel intensity indicates the likelihood of a specific body part being located at that position (see, e.g. [8]). Skeleton joints, on the other hand, provide a sparse representation of body pose, typically using a set of key points (or "joints") representing crucial body parts such as elbows, knees, wrists, and shoulders. These joints are connected by lines to form a skeleton-like structure, representing the person's pose (see, e.g. [9]). For each type of representation, we explore two possible methods in Fig. 3 without introducing additional training parameters to the model, making them adaptable to any VTON model:

- **Concatenate Pose in the Spatial Dimension.** A straightforward approach to incorporating additional conditions is to concatenate them after the reference image: $x_{\text{masked}} = \text{concat}_s([x_m, x_g, x_c]), x_{\text{mask}} = \text{concat}([m, x_\beta, x_\beta]), x_{\text{noisy}} = \text{concat}([x_t, x_\alpha, x_\alpha])$. The final output only utilizes the first third along the spatial dimension.
- **Stitch Pose into the Masked Area.** Since the pose constraint originates from the source image, we can integrate the pose and masked image as a single image using $I_m = I_p \otimes M + I_c \otimes (1 - M)$.

3.3.1. Masking Strategy

In inpainting tasks, the mask plays a crucial role in marking the boundary of the context and conveying the user's intent. For various pose conditioning scenarios, we consider corresponding masking strategies:

- **Fine-grained Mask:** In this scenario, users employ the mask to precisely define the boundary of the inpainting area, indicating how the clothing should be overlaid onto the person. This strategy is primarily used for pose preservation.
- **Bounding Box Mask:** When users aim to adjust the original pose by providing a pose image not derived from the source image, the fine-grained mask may not effectively capture the intended boundary. To offer greater flexibility in model generation for different poses, a bounding box is obtained by finding the minimal rectangular that covers the fine-grained mask. It is used to represent the approximate area being edited.

During training, the fine-grained and bounding box masks are applied alternately to help the model recognize and differentiate between these distinct intentions.

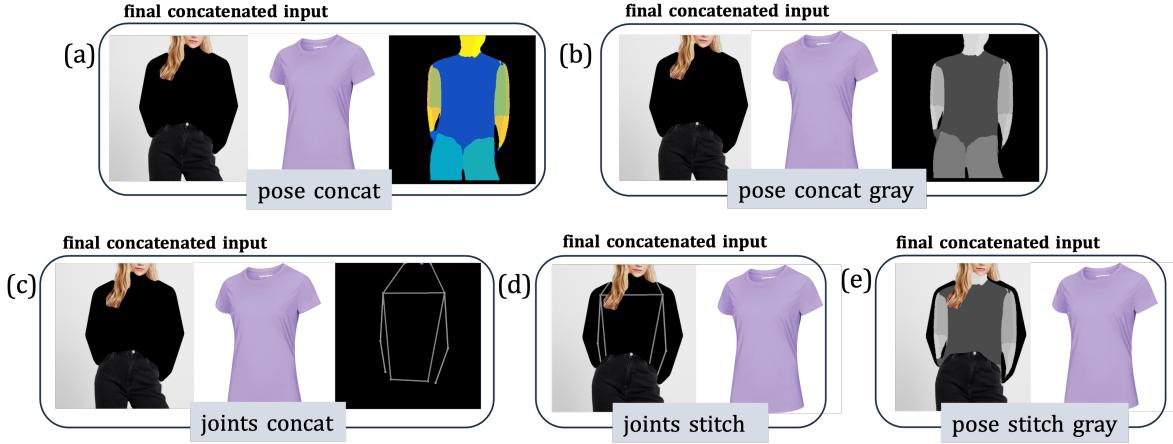


Figure 4. Illustration of different methods for combining pose representation (joints vs. pose map), integration techniques (stitching vs. concatenation), and color modes (color vs. grayscale).



Figure 5. Qualitative comparison of baseline pose-free/pose-less models and our methods for pose preservation.

4. Experiment

4.1. Datasets

We conduct our tests on two publicly available virtual try-on datasets: VITON-HD [10] and DressCode [11]. Each input image is padded and resized to a resolution of 512×512 . VITON-HD contains 2,032 image pairs of upper-body garments in its test split. We utilize the preprocessed masks provided in this dataset. DressCode comprises 5,400 testing pairs across three garment categories (upper-body, lower-body, and full dresses).

4.2. Pose Representations

For pose maps, VITON-HD test set provides preprocessed pose maps directly within the dataset, while for DressCode, the results extracted from a pose map representation model (similar to [8]). We then apply the same color map as VITON-HD to generate pose visualization images. For joint

skeletons, we generate joint points using a joints pose representation model.

4.3. Adding Pose Conditions

To integrate pose conditions into the model inputs, we experiment with two pose representations (skeleton joints and pose maps) and two input formats (concatenation and stitching). Additionally, preliminary studies suggest that color in pose images can interfere with generation quality. Therefore, we test both color and grayscale forms for pose and skeleton images, resulting in five control groups, as illustrated in Fig. 4:

- **Joints Stitch:** Stitching the skeleton joints image into the masked regions.
- **Joints Concat:** Concatenating the skeleton joints image along the spatial dimension, adjacent to the reference image.
- **Pose Concat:** Concatenating the pose map image along

- the spatial dimension, adjacent to the reference image.
- **Pose Concat Gray:** Concatenating the grayscale-converted pose map image along the spatial dimension, adjacent to the reference image.
 - **Pose Stitch Gray:** Stitching the grayscale-converted pose map image into the masked regions.

4.4. Implementation Details

We employ Stable Diffusion v1.5 as our diffusion backbone models [4, 32]. The model is trained with a learning rate of 1e-5 and a batch size of 8. We use DDIM with 25 sampling steps and a guidance scale of 5. Both diffusion models are fine-tuned for 20 epochs on 4 NVIDIA H100 80G GPUs. During training, we use 50% fine-grained masks and 50% bounding box masks.

4.5. Baselines and Evaluation Metrics

We select the latest baseline results from CatVTON [24], StableVTON [19], LaDI-VTON [18], IDM-VTON [20]. To evaluate the model’s output, we employ the widely used evaluation metrics: Structural SIMilarity Index (SSIM) [35] and Learned Perceptual Image Patch Similarity (LPIPS) [36], comparing generated results to ground truth images. In the unpaired setting, where the garment in the person image differs from the input garment image, we utilize Fréchet Inception Distance (FID) [36] and Kernel Inception Distance (KID) [37] as additional metrics. Our implementation follows that in [18].

5. Results

5.1. Quantitative Results

5.1.1. Pose Maps vs Skeleton Joints

Comparing pose concatenation rows with joints and pose maps (see Tab. 1), we observe that pose maps consistently outperform skeleton joints across all metrics. For stitched results, the grayscale pose maps row yield superior outcomes than the joints stitching row. This suggests that pose maps capture richer pose information, which enhances the quality of the generated results.

5.1.2. Concatenating vs Stitching

Within each pose representation, we compare the concatenation and stitching methods. As shown in Tab. 1, stitching outperforms concatenation for both representations. Additionally, converting pose maps to grayscale does not enhance performance in concatenation methods but significantly improves stitching methods. This finding implies that direct concatenation may not fully leverage the pose information. To better integrate the pose map, stitching with grayscale pose maps is preferred.

5.1.3. SD vs SDXL

In Tab. 1, a comparison between the left and right sections shows that SDXL and SD v1.5 achieve comparable results. Observations regarding pose representations and integration methods are consistent across both model architectures.

5.1.4. Comparison with Baseline Results

In Tab. 2, we explore the potential benefits of pose conditioning on VTON performance. Building on the original encoder-free VTON architectures in [1, 34] we stitch the grayscale pose map into the masked area and test both SD v1.5 and SDXL inpainting models as backbones. Integrating the pose image without introducing additional training parameters consistently improves the original results and outperforms other baselines that add an extra channel for the pose image. This comparison demonstrates that stitching grayscale maps can enhance generation results in an effective manner.

5.2. Qualitative Results

5.2.1. Pose Preservation

Fig. 5 compares a pose-less baseline model and our model in terms of pose preservation. When parts of the body are masked (especially limb extremities like hands and feet), pose-less model often misinterprets the pose, and in the second example, even alters the T-shirt to a long-sleeve style. In contrast, our model, guided by pose conditions, more accurately preserves the original pose and retains the clothing style.

5.2.2. Pose Control

In Fig. 6, we illustrate how our model adapts to different poses based on the input pose control signals. We compare our results with IDM-VITON (we use a new pose image as its input during inference to guide generation). Our method effectively generates poses that differ from those in the original source images, offering flexible user-defined poses. While IDM-VITON accurately reflects the target pose, its outputs appear less natural in comparison.

5.2.3. Ablations

Fig. 7 presents the generation results from the five control groups. We observe that pose concatenation introduces unnatural color artifacts, while joint stitching produces reasonable color but tends to overfit within the masked area. The grayscale pose stitching method, in comparison, more accurately captures both the pose and body shape of the person.

Additional comparisons of baseline pose-free/pose-less models and our methods for pose preservation, as well as failure cases of pose-free model are presented in Fig. 8,

5.2.4. Limitations

In Fig. 9, we illustrate a failure case where the input pose image is misaligned with the context outside the bounding



Figure 6. Qualitative comparison of IDM-VTON and our methods for pose-controlled try-on generation.



Figure 7. Qualitative comparison of the five methods for pose integration into the model input.

box, causing the generated results to deviate from the expected pose. This highlights the importance of maintaining consistent depth and positioning of the human figure relative to the original pose to achieve more natural results. Extending the model to handle more diverse poses with variations in position and depth will be explored in future work.

6. Conclusion

In this work, we explore different pose representations and ways to condition pose as the VTON model input. Experimental results show that stitching grayscale pose map into the masked area of the source image achieves the best gen-

eration results and improve the original base model. Meanwhile, it outperforms other baselines that use additional channel to accommodate pose signals. Our method can be flexibly adapted to any VTON models. It is shown to not only preserve the original pose but also generate versatile poses of the original user.

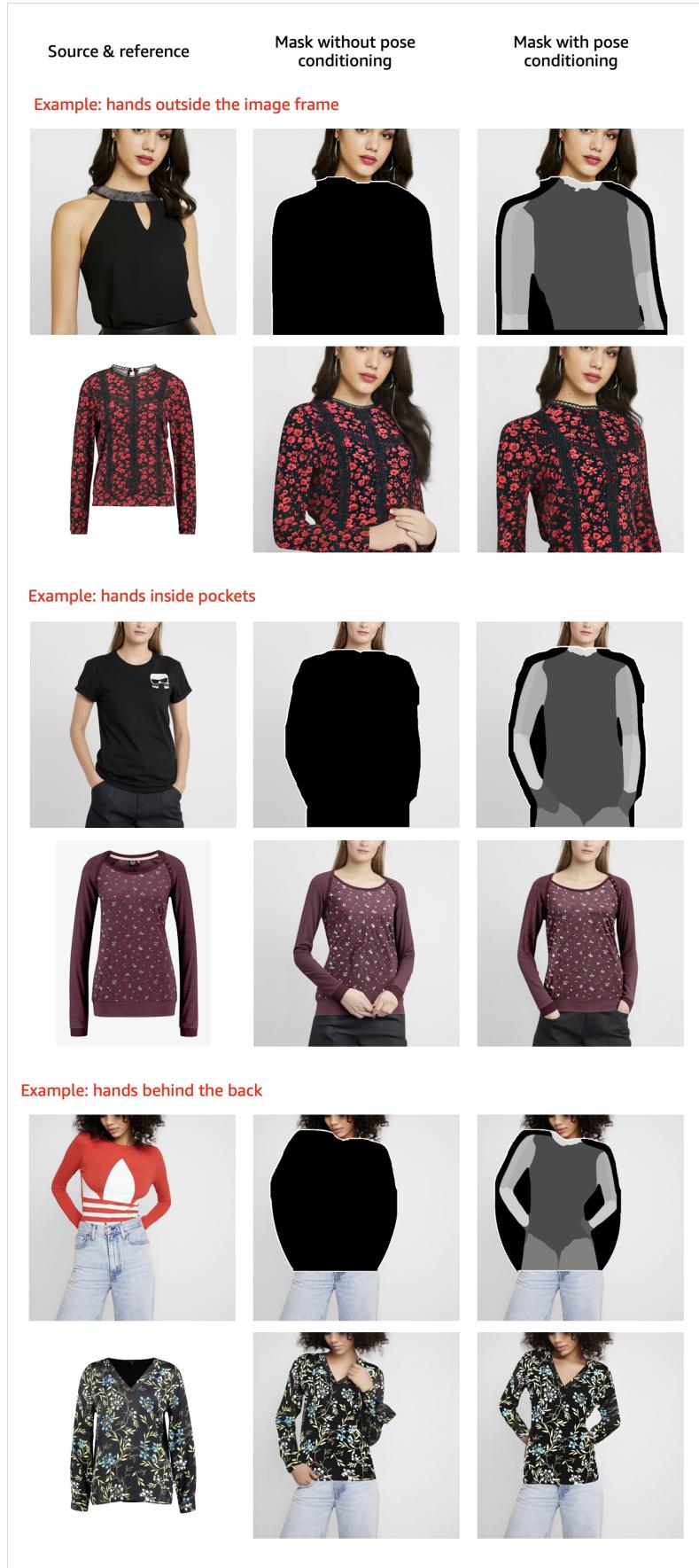


Figure 8. Additional comparisons of baseline pose-free/pose-less models and our methods for pose preservation. We show typical failure cases of pose-free model.



Figure 9. A failure case where the provided pose image is misaligned with the original unmasked context in depth and position.

References

- [1] Qi Li, Shuwen Qiu, Julien Han, Xingzi Xu, Mehmet Saygin Seyfioglu, Kee Kiat Koo, and Karim Bouyarmane. Dit-vton: Diffusion transformer framework for unified multi-category virtual try-on and virtual try-all with integrated image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. AI for Content Creation Workshop. 1, 2, 4, 7
- [2] 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] 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. 1, 2
- [4] 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, 7
- [5] 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.
- [6] 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.
- [7] 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
- [8] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018. 2, 3, 5, 6
- [9] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2, 3, 5
- [10] 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, 6
- [11] 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. 3, 6
- [12] Xiu Li, Michael Kampffmeyer, Xin Dong, Zhenyu Xie, Feida Zhu, Haoye Dong, Xiaodan Liang, et al. Warpdifusion: Efficient diffusion model for high-fidelity virtual try-on. *arXiv preprint arXiv:2312.03667*, 2023. 3
- [13] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7543–7552, 2018. 3
- [14] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 589–604, 2018.
- [15] 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.
- [16] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8485–8493, 2021. 3
- [17] 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] 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. 3, 5, 7
- [19] 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. 3, 5, 7
- [20] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for virtual try-on. *arXiv preprint arXiv:2403.05139*, 2024. 3, 4, 5, 7
- [21] 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
- [22] 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
- [23] 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

- [24] 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
- [25] 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
- [26] 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
- [27] 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
- [28] 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. 3
- [29] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 3
- [30] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 3
- [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
- [32] 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, 7
- [33] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 4
- [34] Xingzi Xu, Qi Li, Shuwen Qiu, Julien Han, and Karim Bouyarmane. Deft-vton: Efficient virtual try-on with consistent generalised h-transform. <https://deft-vton.github.io/>, 2024. 4, 7
- [35] 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
- [36] 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
- [37] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 7