Time-Efficient and Identity-Consistent Virtual Try-On Using A Variant of Altered Diffusion Models

Phuong Dam[1], Jihoon Jeong[1], Anh Tran[2], and Daeyoung Kim[1]

[1] Korea Advanced Institute of Science and Technology (KAIST)
{hoangphuong1211, jihooni, kimd}@kaist.ac.kr
[2] VinAI Research
v.anhtt152@vinai.io

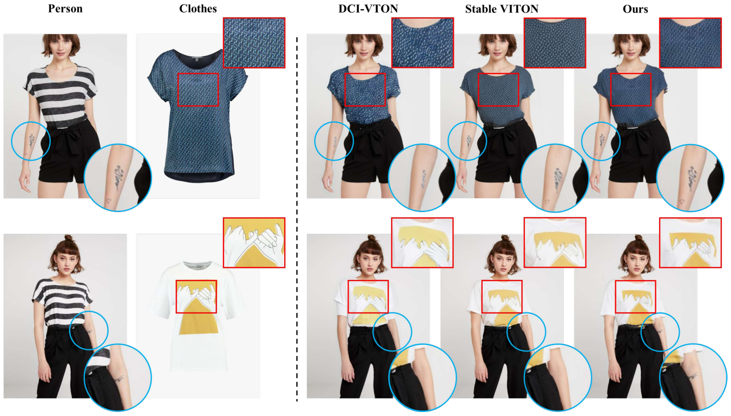


Fig. 1: Visualization for identity preservation and detail preservation compared with DCI-VTON and StableVITON [6, 14]. Both models tend to degrade the texture of clothes, struggle with maintaining symbols on garments, and produce in noticeable artifacts while our approach maintains the fidelity of both garment textures and tattoos.

Abstract. This study discusses the critical issues of Virtual Try-On in contemporary e-commerce and the prospective metaverse, emphasizing the challenges of preserving intricate texture details and distinctive features of the target person and the clothes in various scenarios, such as clothing texture and identity characteristics like tattoos or accessories. In addition to the fidelity of the synthesized images, the efficiency of the synthesis process presents a significant hurdle. Various existing approaches are explored, highlighting the limitations and unresolved aspects, e.g., identity information omission, uncontrollable artifacts, and low synthesis speed. It then proposes a novel diffusion-based solution that addresses garment texture preservation and user identity retention during virtual try-on. The proposed network comprises two primary modules - a warping module aligning clothing with individual features and a try-on module refining the attire and generating missing parts integrated with a mask-aware post-processing technique ensuring the integrity of the individual's identity. It demonstrates impressive results, surpassing the state-of-the-art in speed by nearly 20 times during inference, with superior fidelity in qualitative assessments. Quantitative evaluations confirm comparable performance with the recent SOTA method on the VITON-HD and Dresscode datasets. We named our model **F**ast and **I**dentity **P**reservation **Vi**rtual **T**ry**ON** (FIP-VITON).

Keywords: Time efficiency Virtual Try-on · Identity retention Virtual Try-on · Mask-aware post-processing · Diffusion-based networks

1 Introduction

Virtual Try-On, which involves placing a garment on a particular individual, holds crucial significance in contemporary e-commerce and the prospective metaverse. The key challenge lies in preserving intricate texture details and distinctive features of the target person, such as appearance and pose. Adapting a garment to different body shapes without altering patterns is particularly challenging, especially when body pose or shape varies significantly.

Recent studies based on deep learning techniques have approached these challenges by defining specific body-garment corresponding regions, particularly addressing obstructions [30] or by adding cloth segmentation information [15]. Another approach is taking advantage of the strength of the Diffusion network, combined with Contrastive Language-Image Pretraining (CLIP) [23], which refines post-warping clothing results along with generating missing body parts [6]. Another method uses implicit warping integrated with Diffusion, guided by CLIP-based networks, to overcome this limitation [32]. All the diffusion-based approaches are proven to surpass the traditional flow-based methods in both quantitative and qualitative assessment [6, 32].

Despite strong generative ability, diffusion-based approaches suffer from extended inference times and uncontrollable artifact generation, affecting the user experience and image fidelity. Meanwhile, another equally important challenge besides the garment texture preservation is retaining the user's identifying characteristics during virtual try-on - mentioned in [32]. Therefore, to tackle these issues, we propose a novel diffusion approach that not only effectively preserves both the garment texture and identity information but also achieves an impressive inference speed for this task.

Our network comprises two primary modules - a warping module and a try-on module, integrated with post-processing blocks. The warping module is pivotal in aligning clothing with the individual's features. It considers clothing specifics and person-related information, encompassing key points, dense pose images, and garment type-specific regions of interest (e.g., upper, lower, or full dresses). Subsequently, the try-on module refines the warped attire from the initial module, generating the missing parts of the image. The image then undergoes a conditional post-processing named mask-awareness technique to ensure the fundamental integrity of the individual's identity. Examples of our impressive results compared to those from the SOTA papers [6,14] are depicted in detail in Fig. 1.

In summary, the main contributions of our work are:

- We introduce a novel try-on technique to generate photo-realistic results for diverse scenarios.
- We introduce a novel time-efficient diffusion approach that can adjust and maintain the garment details and generate the missing body parts using sophisticated conditional modules, which effectively guide the model's focus during the generation process to yield satisfying outcomes.
- We introduce a mask-aware post-processing technique that not only preserves the individual's identity details but also improves the overall fidelity of the generated images.

2 Related Work

2.1 Virtual Try-on GAN-based Models

In virtual try-on, achieving high levels of realism and fidelity in garment rendering on digital avatars remains a significant challenge. Traditional deep-learning approaches, primarily utilizing flow-based Generative Adversarial Networks (GANs), have demonstrated notable potential. Most existing virtual try-on methods follow a two-stage process [4,15,30]. The first stage involves a warping module responsible for predicting the appearance flow for the global [4,15], or local parts [30] of the garment to fit the target person's pose. This is followed by a second stage, where a GAN-based generator seamlessly integrates the warped garment into the model. Although this method has shown some effectiveness, its heavy reliance on warping quality in the initial stage often leads to less-than-ideal outcomes. This is particularly evident in the realism of garment-skin boundaries and the overall try-on effect in the garment area.

Despite their widespread use, these methods have not seen significant innovation to overcome these specific challenges. *Therefore, we proposed an approach that aims to push the boundaries of the field by investigating the use of diffusion-based generators.*

2.2 Virtual Try-on Diffusion Models

Diffusion models have recently emerged as formidable rivals to GANs in image generation, excelling in producing high-fidelity conditional images. Their application "diffuses" a diverse range of tasks, including text-to-image generation, image-to-image translation, and image editing. Notably, in the context of virtual try-on, diffusion models offer a promising solution by treating the task as a form of image editing conditioned on a specific garment and a full-body image.

A notable diffusion-based approach is presented in TryonDiffusion [32], which leverages a cross-attention mechanism and integrates an implicit warping algorithm with a try-on module. While this method showcases potential in virtual try-on applications, *it struggles to retain textural details in the final output.*

Other strategies combine the explicit warping module from GAN-based methods with a diffusion model to merge warped clothing and person images. LaDI-VTON [20], DCI-VTON [6], and StableVITON [14] leverage the capabilities of Stable Diffusion to preserve the texture and details of in-shop garments, achieving high-quality images to a certain level. However, these methods often create unintended artifacts in the final images, which remain challenging to control. Despite the outstanding generation quality of the diffusion approach, *the models' inference times are excessively long.* The major reason lies in the *large number of denoising steps during the inference process.*

2.3 Diffusion Model Speed Up Techniques

Recent research has focused on accelerating the inference time of Diffusion model networks. Most of these methods are based on types of distillation techniques, such as [18,19,25]. Meanwhile, in specific terms of virtual try-on, no generalized approaches for multiple timestep diffusion have been distilled. Although [25] offers a direct training method that allows generating images with a single timestep, there is insufficient evidence that this method works well in a conditional high-resolution diffusion model, leaving space for future research. Furthermore, several studies applied a multimodal conditional GAN to reduce the number of timesteps to 4 or 2 [22,29], speeding up the inference process of the diffusion model. *Inspired by leveraging multimodal GAN* [22,29], we opt to experiment with a single-step diffusion-based approach employing a fixed noise level based on Denoising Diffusion Implicit Models (DDIM) [24].

3 Methodology

Addressing these challenges, we propose a novel approach to enhance artifact control. Our method involves developing a diffusion model that not only competes with state-of-the-art(SOTA) models' performance but also ensures time efficiency. This is achieved by incorporating *(un)conditional mask-aware techniques* and *a modified adding noise algorithm reducing the number of time steps to one*, thereby offering a practical and effective solution for virtual try-on applications. The mask-aware techniques are particularly crucial in attaining a realistic wearing effect, as they directly address the common issue of unnatural transitions between clothing and skin, and preserve individual's identities.

In the virtual try-on task, given an image of a person I_p and an image of a garment I_g, we want to obtain the try-on image that portrays the person wearing the garment. The overall architecture of our approach is depicted in Figure 2. We first pre-process the person image to obtain reference information, including human parsing, 2D pose key points J_p, and dense pose I_{dp}. Following this, our wrapping module utilizes the reference information and produces predicted multi-scale warped-based parsing $\{S_i\}$ including both the warped cloth mask S_m^i and the segmentation of visible body parts S_{bp}^i, along with multi-scale appearance flows $\{f_i\}$. At the highest resolution ($i = 1$), these outputs are integrated

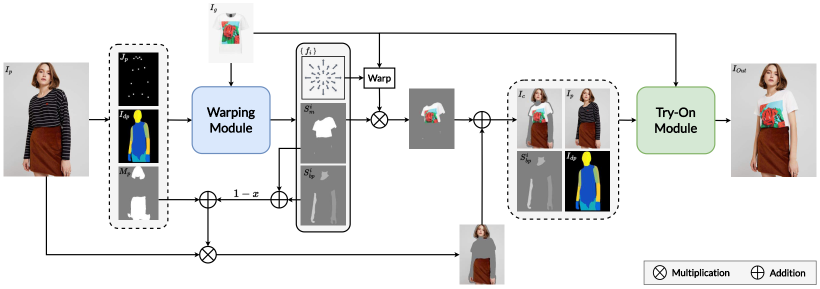


Fig. 2: Overall Generation Pipeline.

with the garment and person images, in addition to preserved person information (M_p), to construct the conditional input (I_c). The conditional input I_c in conjunction with dense pose I_{dp}, predicted body part segmentation S^i_{bp}, garment image I_g, and person image I_p are fed into the try-on module to produce the final image I_{Out}.

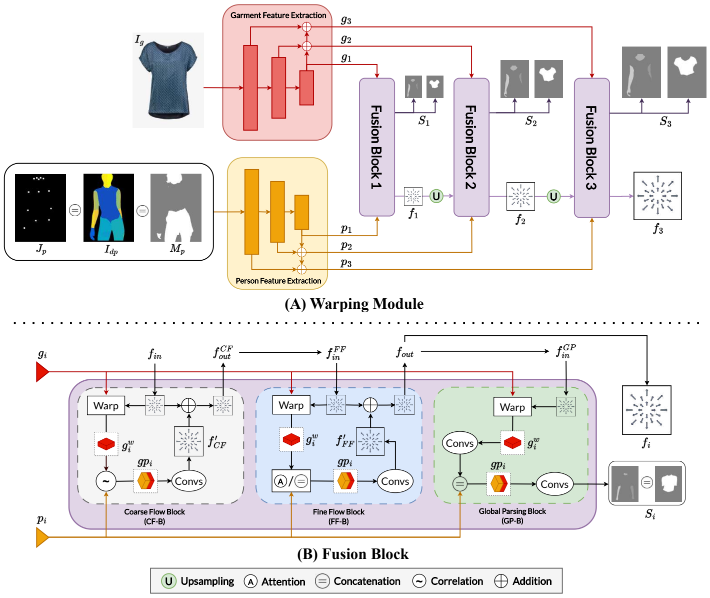
3.1 Preprocessing

The procedure begins with generating human parsing maps from the person image (I_p) by employing advanced human parsing methods [16]. Then, we apply 2D key points [3], and dense pose estimation [7]; their outputs are denoted as J_p and I_{dp}, respectively. Subsequently, depending on the garment type, we *identify and exclude mutable sections* based on the human parsing map. Specifically, this mechanism is that if the garment is upper types, the upper regions of the body are omitted, which works the same as the lower types and dresses – the full body types. The remaining segments are combined to create the preserved mask (M_p).

3.2 Warping Module

In this section, we detail the structure of the Warping Module. As illustrated in Figure 3, our warping module draws inspiration from the flow estimation pipeline found in [5, 8, 9, 15, 30]. It comprises two pyramid feature extraction blocks, "Garment Feature Extraction" and "Person Feature Extraction" modules; cascade flow estimation. Our distinct contributions in this module will be elucidated below.

Pyramid Feature Extraction. Our warping module leverages two Feature Pyramid Networks (FPN) [17] for extracting N multi-scale person and garment features. The person feature extraction block receives inputs from the 2D human pose keypoints map J_p, dense pose I_{dp}, and the preserving region mask M_p. Conversely, the garment feature extraction block exclusively takes inputs from the intact in-shop garment I_g. Notably, we have no change compared to the FPN [17] used in [30] except the number of scales (N) varies on input resolution.



(A) Warping Module

(B) Fusion Block

Fig. 3: Warping Module structure. It is crucial to highlight that our model extracts six or seven multi-scale features, depending on the input resolution (i.e., $N = 6$ or 7). For brevity, the number of scales depicted in this figure is limited to three ($N = 3$).

Cascade Flow Estimation. Inspired by established methods [5, 8, 9, 15, 30], instead of estimating the local flows for certain parts of the cloth as in [30], we target the direct prediction of global flow for warping intact garments. Meanwhile, our warping module, depicted in Figure 3, adopts concepts from [30] with internal enhancement. *Notably, our enhancement introduces cross-attention for additional feature integration, elevating the quality of the warping outcome and thereby enhancing the overall model performance.*

Specifically, our module incorporates N Fusion blocks designed to handle multi-scale flow maps and human parsing predictions. Each Fusion block is composed of a Coarse Flow Block (CF-B), a Fine Flow Block (FF-B), and a Global Parsing Block (GP-B), depicted in gray, blue, and green, respectively, Fig. 3(B).

In CF-B, the garment feature g_i undergoes warping with incoming flow f_{in} to produce g_i^w. The correlation operation from FlowNet2 [11] integrates it with the person feature p_i, and subsequent convolution layers estimate the corresponding flow f'_{CF}. The refined coarse flows f_{out}^{CF} result from adding f'_{CF} to f_{in}. FF-B, sharing CF-B's architecture, treats f_{out}^{CF} as the input flow f_{in}^{FF}. Diverging from CF-B, FF-B opts for multi-head cross-attention instead of correlation, using scaled dot-product attention [26]:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V \tag{1}$$

where $Q \in \mathbb{R}^{M\times d}, K \in \mathbb{R}^{N\times d}, V \in \mathbb{R}^{N\times d}$ represent stacked vectors of query, key, and value, respectively. M is the number of query vectors, N is the number of key and value vectors, and d is the dimension of the vector. In our setup, Q represents the flattened feature of the warped garment g_i^w, while K and V correspond to the flattened features of the person p_i. The dot-product-based attention map $\frac{QK^T}{\sqrt{d}}$ serves as an additional feature indicating the similarity between the person and the warped garment. Moreover, we restrict the application of multi-head cross attention to the feature resolution below 64x48 for parameter efficiency, switching to concatenation at larger resolutions. The output of this operation is directed to a group convolution block for estimating flow f'_{FF}, which is added to f_{in}^{FF} to yield the fine flow f_{out}.

Within the Global Parsing Block (GP-B), leveraging the enhanced flow f_{in}^{GP}, the garment feature g_i is warped. The newly warped feature g_i^w undergoes fusion with the incoming person feature p_i through convolution operations. The concatenated feature gp_i is then processed by convolutional layers to estimate the global parsing S_i, covering background, cloth, left/right arms, center body parts (including neck and belly), and left/right legs.

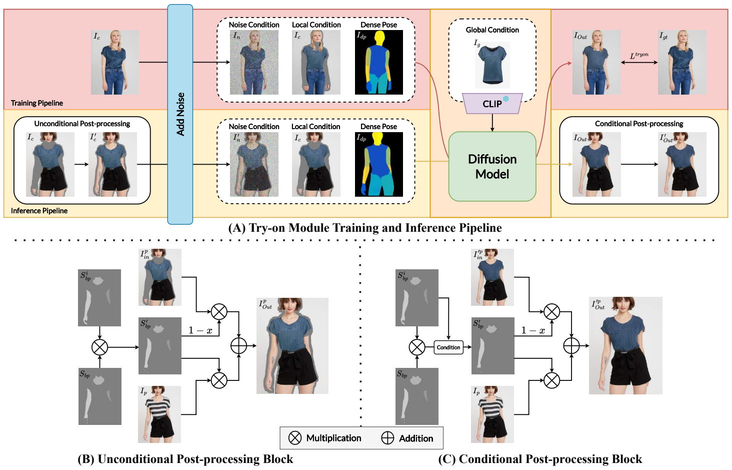
Objective Function. Similar to numerous prior studies employing flow-based warping models [5, 8, 9, 15, 30], our warping model follows a similar structure, incorporating a combination of various loss functions. In the training of the warping module, we utilized ℓ_1 loss - L_1^w and perceptual loss [12] - L_{per}^w for the warped result. Additionally, pixel-wise cross-entropy loss L_{ce}, and ℓ_1 loss L_m^w are applied to the entire estimated parsing. Adversarial loss [13] - L_{adv}^w is employed for both overall parsing and the warped result. Given the appearance flow's high degree of freedom, we also include total-variation (TV) loss L_{TV} as proposed in [6], effectively addressing the smoothness of the final warping result. In alignment with the approach outlined in [5], we augment a second-order smooth constraint L_{sec}. The total loss function for our warping module can be formulated as:

$$L^w = L_1^W + \alpha_{per}^w L_{per}^w + \alpha_{ce} L_{ce} + \alpha_m^w L_m^w + \alpha_{adv}^w L_{adv}^w + \alpha_{TV} L_{TV} + \alpha_{sec} L_{sec}, \quad (2)$$

with α_{per}^w, α_{ce}, α_m^w, α_{adv}^w, α_{TV}, and α_{sec} are hyper-parameters controlling the contribution of each loss term to the overall loss.

3.3 Try-on Module

Despite the robust stability exhibited by traditional vanilla diffusion models, they suffer from a notable drawback — prolonged inference times, which can hinder practical applications. *However, our variance approach addresses this issue without compromising the model's generative performance.* As indicated in the overview of our strategy in Fig.2, we intend to apply a variant of the modified diffusion model to refine the coarse synthesis result. Following the acquisition of the warping and parsing result, the warped garment is incorporated into the person's image, alongside the preserved elements and the original background,



(A) Try-on Module Training and Inference Pipeline

(B) Unconditional Post-processing Block

(C) Conditional Post-processing Block

Fig. 4: Try-on module training, inference strategy, and details. S_{bp} is the body part segmentation extracted from I_p.

to form the conditional image I_c. This image I_c serves as the local condition for the try-on diffusion module. Moreover, the global conditions are derived from the target garment image I_g using the feature representations extracted by a frozen, pre-trained CLIP [23] image encoder. All are illustrated in Fig. 4(A).

Training pipeline. During the training process, we work with pair-setting; a noise level is directly added to the ground truth image I_{gt} to create a conditional noise input I_n. Subsequently, the local condition image I_c mentioned earlier is concatenated with the dense pose image I_{dp} - dense pose condition, and the conditional noise I_n. This concatenated input is then fed into the model. The output from the model is then merged with the static components, which include the overlapping areas of the predicted and actual backgrounds as well as the unaltered segment M_p. *This highlights our try-on model's focus on generating missing image parts and refining clothing details.*

The objective function is then applied to the integrated output and the ground truth image. Our model leverages global and local conditions in the optimization process to generate the corresponding inferred person image. In contrast to prior studies related to the diffusion model, where the amount of noise added to the image is adjusted through a time variable, in this study, we use a fixed amount of noise, rendering the time variable redundant during the training process. *At this point, our approach could be considered a single-step diffusion model.* The adding noise process follows the equation:

$$z = z_0 + \alpha_n \epsilon, \tag{3}$$

where α_n is the noise ratio and $\epsilon \sim \mathcal{N}(0, I)$. Therefore, our framework is built upon the baseline model of the DDIM study [24], with the temporal embedding block supplanted by the CLIP feature.

Inference pipeline. In the absence of the ground truth image during inference, it becomes imperative to generate an alternative conditional input to facilitate the inference pipeline. In the inference process, the local condition I_c is constructed and sent to a post-processing block after the warping module. Specifically, the output of the first post-processing block I'_c is created by combining the local condition image I_c with the remaining body parts extracted from the ground truth image I_p. This combined image is then made noisy to obtain another noise-conditioned image I'_n. Following a similar diffusion process as outlined in the training phase, the output image I_{Out} is advanced into another post-processing block, culminating in the final result I'_{Out}. This sequence ensures the generation of a refined output image, compensating for the absence of ground truth data through strategic conditioning and processing stages.

(Un)conditional Post-processing Block. The post-processing technique, termed (un)conditional mask-aware, operates throughout the entire inference pipeline, ensuring the enhancement of synthesized images with a focus on artifact reduction and detail preservation. *The first post-processing technique, unconditional post-processing, aims to enhance the condition image, improving the final synthesized results. Meanwhile, the second technique, conditional post-processing, focuses on minimizing artifacts and preserving identity details.*

The unconditional post-processing block involves obtaining body segmentation masks S^i_{bp} here $i = 1$ denoting the largest segmentation map extracted from the warping module and corresponding body masks S_{bp} from the reference person image. The overlapping mask S'_{bp} is used to extract unchanged elements from I_p, which are then added to I_c to get the coarse try-on image I'_c.

The conditional post-processing block is used to re-apply the unchanged parts from the input image I_p to the refined try-on image I_{Out}, correcting any undesirable modifications caused by the diffusion model. It has the same structure as the unconditional one, except that it is only applied when the overlapping ratio between S_{bp} and S^i_{bp} where $i = 1$, is greater than a threshold. Empirically, we found this threshold R_{overlap} needs to be greater than 0.8. This experiment is detailed in the Appendix, providing the applied rate of each specific threshold.

$$R_{\text{overlap}} = \frac{\left|S^{(1)}_{bp} \bigcap S_{bp}\right|}{S^{(1)}_{bp}} = \frac{S'_{bp}}{S^{(1)}_{bp}} \tag{4}$$

Notably, we apply this to every single parsing part, including the left/right hand, left/right leg, and center body part (neck and belly), shown in Fig. 4(B),(C).

Objective function. In terms of optimization processing for the try-on module, we utilize 2 reconstruction losses which are ℓ_1 loss - L^{tryon}_1 and VGG perceptual loss [12] - L^{tryon}_{per}. Furthermore, an adversarial loss from [13] is also used for better quality results - L^{tryon}_{adv}. The total loss for the try-on module can be formulated:

$$L^{\text{tryon}} = L^{\text{tryon}}_1 + \alpha^{\text{tryon}}_{per} L^{\text{tryon}}_{per} + \alpha^{\text{tryon}}_{adv} L^{\text{tryon}}_{adv} \tag{5}$$

where, $\alpha^{\text{tryon}}_{per}$ and $\alpha^{\text{tryon}}_{adv}$ are the balance coefficients.

4 Experiments

4.1 Experiments Setting

Dataset. Our experiments primarily use the VITON-HD dataset [4], comprising 13,679 frontal-view woman and upper clothes image pairs at 1024x768 resolution. Following prior work [4,15], we split the dataset into a training set of 11,647 pairs and a test set of 2,032 pairs. Experiments are conducted at various resolutions, and we also assess the model on the DressCode dataset [21] for added complexity. Much like VITON-HD [4], the DressCode dataset [21] is a repository of high-quality try-on data pairs, comprising three distinct sub-datasets: dresses, upper-body, and lower-body. In total, the dataset encompasses 53,795 image pairs, distributed across 15,366 pairs for upper-body attire, 8,951 pairs for lower-body clothing, and 29,478 pairs for dresses. To maintain consistency, we apply the same human parsing and key points pose estimation methods used on the DressCode [21] to the VITON-HD [4].

Evaluation Metrics. In Virtual Tryon evaluations, we consider paired and unpaired settings. Paired assessments use Structural Similarity Index Measure (SSIM) [28] and Learned Perceptual Image Patch Similarity (LPIPS) [31] for image reconstruction, while unpaired settings employ Frechet Inception Distance (FID) [10] and Kernel Inception Distance (KID) [2] to measure the model's ability to generate new images with changed clothing. Specifically, our evaluation metrics for VITON-HD [4] are all above. Meanwhile, all experiments conducted on the DressCode dataset [21] are executed at a resolution of 512 × 384, and the evaluation metrics only include LPIPS [31], SSIM [28], and FID [10]. In addition, we also measure the speed of synthesizing at the 512 x 384 resolution (T(s)) of the VITON-HD dataset. For a fair comparison in terms of inference time, we ensure consistent configuration settings, employing a single RTX 4090 with a batch size of 4. The model's inference time is computed by averaging the time taken for end-to-end inference over the entire testing set, repeated 10 times.

4.2 Quantitative Evaluation

In our comparative analysis with existing virtual try-on methods on VITON HD dataset [4] and DressCode dataset [21], including CP-VTON [27], VITON-HD [4], FS-VTON [9], SDAFN [1], PF-AFN [5], HR-VTON [15], GP-VTON [30], LaDI-VITON [20], StableVITON [14], and the current state-of-the-art (SOTA) DCI-VTON [6], our method showcases competitive performance across various metrics (Table 1). In terms of VITON HD [4], our model demonstrates compatibility with this SOTA method while DCI-VTON outperforms previous studies in all evaluated categories. Notably, our approach excels in certain aspects while trailing in others, both in paired and unpaired settings. A significant advantage of our model lies in its remarkable inference speed. As detailed in Table 1, our model achieves the best inference times, which is more than 17.43 times faster than DCI-VTON in 512x384 resolution, 1.01s of ours compared to 17.60s of

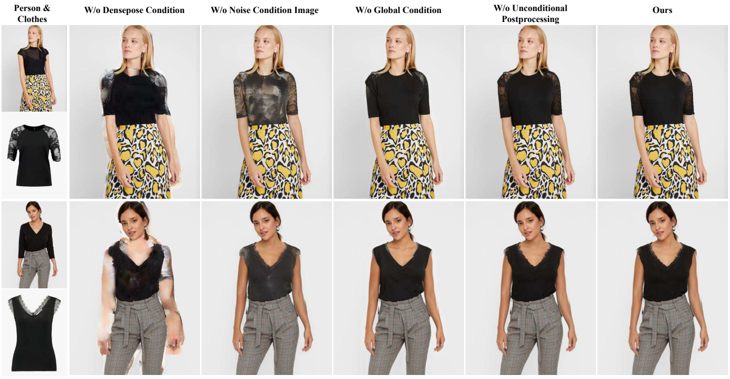


Fig. 5: Visual condition effectiveness ablation studies for the Tryon module in our approach. Note that all ablation studies are applied to conditional post-processing. Please zoom in for better visualization.

DCI-VTON. Besides, our method also demonstrates superior performance that is compatible with the DCI-VTON in almost all three subsets of DressCode [21].

4.3 Ablation Study

By taking 512 x 384 resolution on the VITON-HD dataset as the basic setting, we conduct ablation studies to validate the effectiveness of several components in our networks, and the results are shown in Table 2.

Condition Effectiveness - Tryon Module. In this ablation study, detailed in Table 2, we systematically assess the components of our Tryon module. First, we explore the impact of removing the unconditional post-processing block. In this experiment, the local condition image I_c instead of I'_c is used to make the noise conditions image, we find it slightly degrades the model quality to imply that unconditional post-processing is still important. Second, we investigate the

Table 1: Quantitative comparison with baselines on VITON-HD [4] and DressCode [21]. We multiply KID by 100 for better comparison. T(s)↓ is the inference times. **Bold** and underline denote the best and the second best.

Method	VITON-HD									DressCode - (512 x 384)								
	256 x 192				512 x 384					Upper			Lower			Dress		
	LPIPS↓	SSIM↑	FID↓	KID↓	LPIPS↓	SSIM↑	FID↓	KID↓	T(s)↓	LPIPS↓	SSIM↑	FID↓	LPIPS↓	SSIM↑	FID↓	LPIPS↓	SSIM↑	FID↓
CP-VTON [27]	0.089	0.739	30.11	2.034	0.141	0.791	30.25	4.012	-	-	-	-	-	-	-	-	-	-
VITON-HD [4]	0.084	0.811	16.36	0.871	0.076	0.843	11.64	0.3	**0.64**	-	-	-	-	-	-	-	-	-
FS-VTON [9]	-	-	-	-	-	-	-	-	-	0.0376	0.9457	13.16	0.0438	0.9381	17.99	0.0745	**0.8876**	13.87
SDAFN [1]	-	-	-	-	-	-	-	-	-	0.0484	0.9379	12.61	0.0549	0.9317	16.05	0.0852	0.8776	11.8
PF-AFN [5]	0.089	0.863	11.49	0.319	0.082	0.858	11.3	0.283	-	0.038	0.9454	14.32	0.0445	0.9378	18.32	0.0758	0.8869	13.59
HR-VTON [15]	0.062	0.864	9.38	0.153	0.061	0.878	9.9	0.188	1.39	0.0635	0.9252	16.86	0.811	0.9119	22.81	0.1132	0.8642	16.12
GP-VTON [30]	-	-	-	-	0.08	0.894	9.2	-	-	0.0359	0.9479	11.89	0.042	0.9405	16.07	0.0729	0.8866	12.26
LaDI-VITON [20]	-	-	-	-	0.0986	0.858	12.31	0.567	8.27	0.0654	0.9129	16.18	0.0603	0.9076	16.31	0.1079	0.852	15.80
StableVITON [14]	-	-	-	-	0.073	0.888	8.58	0.073	20.58	0.0388	0.937	**9.94**	-	-	-	-	-	-
DCI-VTON [6]	**0.049**	0.906	8.02	**0.058**	**0.043**	0.896	**8.09**	**0.028**	17.60	**0.0301**	-	10.82	**0.0348**	-	**12.41**	**0.0681**	-	12.25
FIP-VITON (Ours)	0.056	**0.909**	**7.53**	0.07	0.067	**0.909**	8.43	0.066	1.01	0.0357	**0.9495**	10.4	0.0417	**0.9413**	12.69	0.0727	0.886	**11.2**

role of the generated noise condition (I'_n) by substituting it with regular Gaussian noise at the same level. Meanwhile, in the Global Condition effectiveness assessment, we just replaced the CLIP-embedding vector with the same shape zeros vector. In terms of dense pose conditions experiment, we do the same as the global condition case. Our findings underscore the significant influence of both the dense pose condition and the noise condition image, with the former being the most impactful, followed by the latter. Additionally, the global condition - the CLIP-based embedding module shows its importance as the third most influential block. Intriguingly, the removal of the first post-processing block in the Tryon module exhibits minimal impact on model quality. These results are visually demonstrated in Figure 5, where removal of the densepose condition (*w/o Densepose Condition*) destroys body part structures and clothes, while the exclusion of the noise condition image results in pure noise (*w/o Noise Condition Image*), undermining clothing texture. The model's failure to recognize and retain specific cloth textures, evident in the absence of the global condition (*w/o Global Condition*) - the transparent sleeves or the frill of the cloth, highlights its significance. Interestingly, the decision not to apply unconditional post-processing results in only a slight reduction in output detail (*w/o Unconditional Post-processing*).

Additional Ablation Study. Since the number of pages is limited, additional ablation experiment parts are shown in the Appendix. We answer the question *"Does the advantage come from the modification of the Warping Module or come from the stronger prior?"*, *"What is the trade-off between multiple and single timestep diffusion?"*, and *"What is the impact of modification in the Warping Module?"*. Furthermore, there is also an ablation experiment to prove the ability of *plug-and-play* mask-aware postprocessing block.

4.4 Qualitative Evaluation

In Fig. 6, we showcase composite images from various methods on the VITON-HD Dataset at 512 x 384 resolution. Our approach demonstrates superior performance to StableVITON, and DCI-VTON, the current SOTA results on the VITON-HD dataset regarding detail features. Our architecture generates highly

Table 2: Condition effectiveness - ablation study on VITON-HD (512x384). We multiply KID by 100 for better comparison.

Method	**VITON-HD – 512 x 384**			
	LPIPS↓	SSIM↑	FID↓	KID↓
w/o Unconditional Postprocessing	0.0706	0.9045	8.52	0.074
w/o Noise Condition Image	0.0923	0.8774	11.00	0.21
w/o Global Condition	0.0759	0.9036	9.11	0.11
w/o Densepose Condition	0.1725	0.8320	24.49	1.43
Ours	**0.0675**	**0.9091**	**8.43**	**0.066**

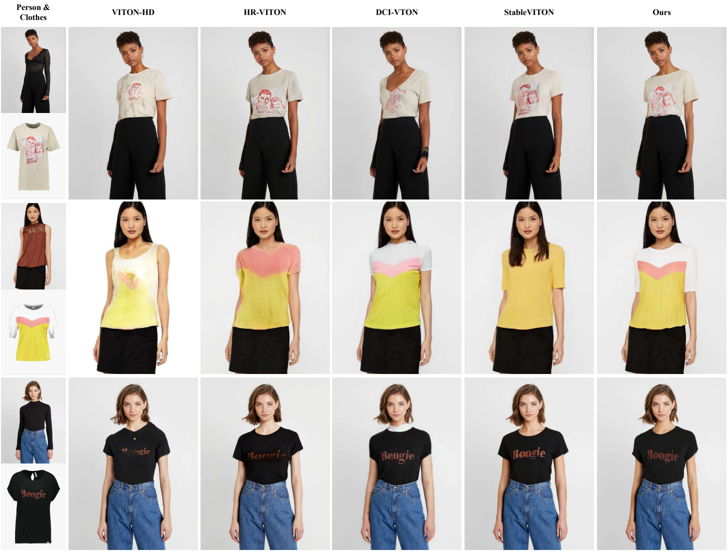


Fig. 6: Qualitative comparison with baseline in VITON-HD Dataset [4] at 512 x 384 resolutions. Please zoom in for better visualization.

realistic images, surpassing previous studies like VITON-HD and HR-VTON, particularly excelling in costume details compared to DCI-VTON and StableVITON. Notably, our model effectively handles challenging cases such as complicated symbols and characters on the clothes the first and third examples, where DCI-VTON and StableVITON fall short. Moreover, in these cases, DCI-VTON introduces artifacts (neck part of the DCI-VTON result), while our approach maintains realism without these artifacts. Furthermore, the second row of Fig. 6 showcases our outperforming in quality in both the scale and reality of the clothes compared with StableVITON and DCI-VTON. The additional qualitative results are shown in the Appendix.

Our research not only preserves outfit details and minimizes artifacts but also addresses the crucial issue of retaining identity information. Leveraging our conditional mask-awareness post-processing technique, we successfully address this concern. Fig. 7 illustrates the efficacy of our post-processing, comparing results before and after its application alongside the outcomes of the SOTA paper under the same conditions. Our technique effectively preserves identity information, demonstrated in cases where DCI-VTON fails to retain arm tattoos (Fig. 1). DCI-VTON introduces artifacts in the neck area (strange necklaces) and the lower-cloth context, compromising image authenticity (1^{st}, 2^{nd} rows). Additionally, DCI-VTON and StableVITON struggle to maintain outfit details, 2^{nd} and 3^{rd} samples in Fig. 7. Our research ensures the maximum retention of identity information for the inference person without compromising output authenticity. The additional qualitative results are shown in the Appendix.

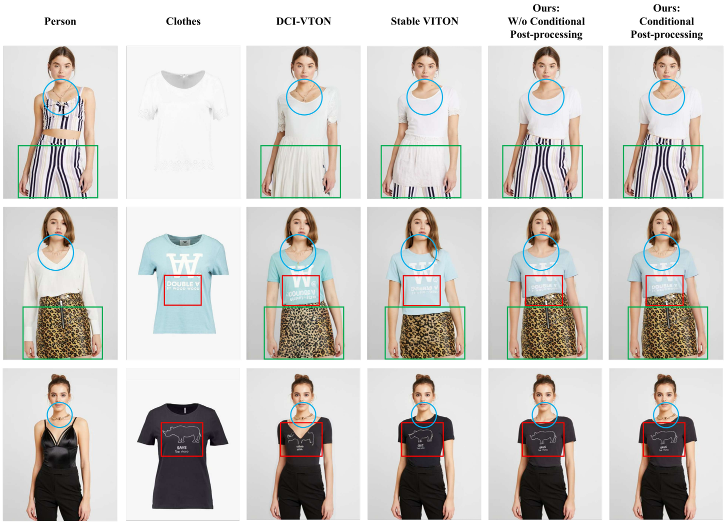

Person
Clothes
DCI-VTON
Stable VITON
Ours: W/o Conditional Post-processing
Ours: Conditional Post-processing

Fig. 7: Example for the identity preservation on VTON-HD Dataset (512 x 384). The blue circle represents the important parts that need to be preserved, the green rectangle is for the wrong context compared to our approach, and the red rectangle is for the detailed part of the clothes that can not be retained compared to our approach. Please zoom in for better visualization.

5 Conclusions

In this study, we have addressed the challenges of Virtual Try-On technology, by introducing a novel diffusion-based approach that adeptly preserves garment texture and user identity. Our integrated system, comprising a warping module and a try-on module, enhanced with a mask-awareness post-processing technique, significantly outperforms existing methods in inference speed, being over 17.43 times faster than the current state-of-the-art, while maintaining superior fidelity in output. This dual focus on efficiency and detail preservation marks a substantial advancement in the field. However, it is important to acknowledge a limitation in our approach: the necessity for an elaborate post-processing process. While crucial for ensuring the integrity of the individual's identity and the garment's texture, this additional step adds complexity to the overall system. Despite this, the proposed method presents a promising solution for real-world applications, offering a more seamless and accurate virtual try-on experience. Future work could aim to streamline this post-processing phase, further enhancing the system's efficiency and applicability in diverse scenarios, including a wider range of clothing styles and body types.

Acknowledgments

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program(IITP-2022-2020-0-01489) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation); and the Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korean government (MSIP) (No. 2022-0-00407).

References

1. 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. pp. 409–425. Springer (2022)
2. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. arXiv preprint arXiv:1801.01401 (2018)
3. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7291–7299 (2017)
4. 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. pp. 14131–14140 (2021)
5. 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. pp. 8485–8493 (2021)
6. 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. pp. 7599–7607 (2023)
7. Güler, R.A., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7297–7306 (2018)
8. Han, X., Hu, X., Huang, W., Scott, M.R.: Clothflow: A flow-based model for clothed person generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10471–10480 (2019)
9. 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. pp. 3470–3479 (2022)
10. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
11. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2462–2470 (2017)
12. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. pp. 694–711. Springer (2016)

13. Jolicoeur-Martineau, A.: The relativistic discriminator: a key element missing from standard gan. arXiv preprint arXiv:1807.00734 (2018)
14. Kim, J., Gu, G., Park, M., Park, S., Choo, J.: Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on (2024)
15. 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. pp. 204–219. Springer (2022)
16. Li, P., Xu, Y., Wei, Y., Yang, Y.: Self-correction for human parsing. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(6), 3260–3271 (2020)
17. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
18. Luo, S., Tan, Y., Patil, S., Gu, D., von Platen, P., Passos, A., Huang, L., Li, J., Zhao, H.: Lcm-lora: A universal stable-diffusion acceleration module. arXiv preprint arXiv:2311.05556 (2023)
19. Meng, C., Rombach, R., Gao, R., Kingma, D., Ermon, S., Ho, J., Salimans, T.: On distillation of guided diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14297–14306 (2023)
20. 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 31st ACM International Conference on Multimedia. pp. 8580–8589 (2023)
21. Morelli, D., Fincato, M., Cornia, M., Landi, F., Cesari, F., Cucchiara, R.: Dress code: high-resolution multi-category virtual try-on. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2231–2235 (2022)
22. Phung, H., Dao, Q., Tran, A.: Wavelet diffusion models are fast and scalable image generators. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10199–10208 (2023)
23. 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. pp. 8748–8763. PMLR (2021)
24. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
25. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models. arXiv preprint arXiv:2303.01469 (2023)
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
27. 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). pp. 589–604 (2018)
28. 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 **13**(4), 600–612 (2004)
29. Xiao, Z., Kreis, K., Vahdat, A.: Tackling the generative learning trilemma with denoising diffusion gans. arXiv preprint arXiv:2112.07804 (2021)
30. 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. pp. 23550–23559 (2023)

31. 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. pp. 586–595 (2018)
32. 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. pp. 4606–4615 (2023)