This ICCV Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

UP-VTON: A Unified Virtual Try-On Framework Supporting Mask, Mask-Free, and Prompt-Driven Guidance

Youngjoo Jo[1,2] Minho Park[1,2] Dong-oh Kang[1]
[1]ETRI [2]KAIST
{run.youngjoo, roger618, dongoh}@etri.re.kr

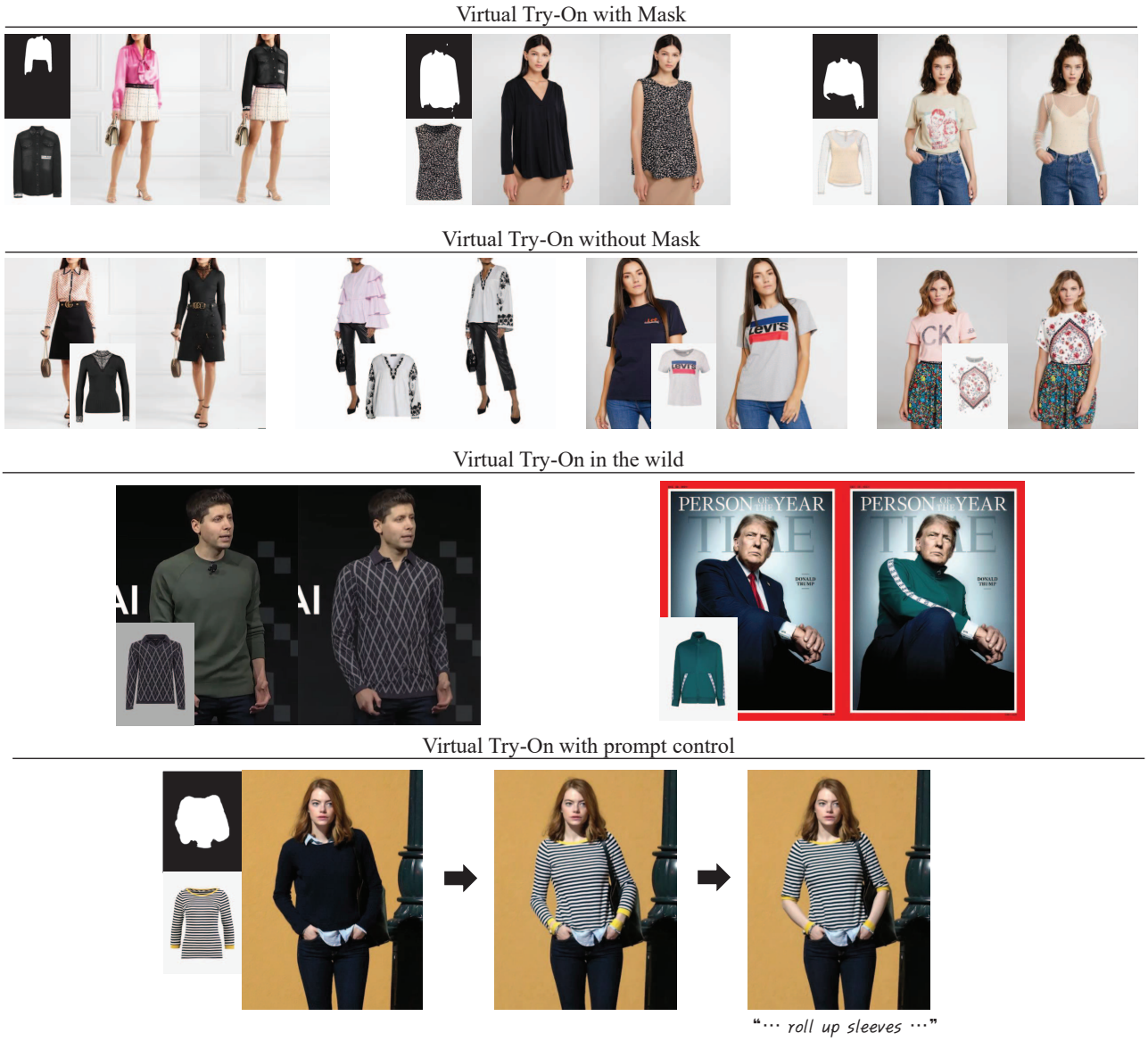

Figure 1. Virtual try-on results generated by UP-VTON. All samples are produced by a single model trained exclusively on the VITON-HD [3] and DressCode [21] datasets. Results on wild (in-the-wild) data are obtained using the model trained solely on the DressCode dataset. All results are generated using a single unified model, which consistently produces high-quality outputs irrespective of the presence or absence of a mask. Furthermore, we demonstrate the model's capability for controllable style generation via prompt-based conditioning.

Abstract

Image-based virtual try-on (VTON) aims to synthesize realistic images of a person wearing a target garment. While recent advances in image generation have improved visual quality, existing methods are typically categorized as either mask-based or mask-free. Mask-based approaches rely on clothing masks to localize garment regions but often cause artifacts and identity distortion. Mask-free methods eliminate this dependency but can suffer from hallucinations and poor garment-person alignment.

We argue that users should be able to control the use and extent of garment masks, as rigid assumptions hinder flexibility and fine-grained editing. Moreover, many prior works require additional modalities—such as keypoints or DensePose—which complicate the pipeline and increase annotation costs.

To overcome these limitations, we propose UP-VTON, a unified virtual try-on framework that performs robustly with or without garment masks and supports prompt-based controllable generation. Our approach introduces triptych prompting, a hybrid inpainting strategy guided by reference images, text prompts, and visual cues. Without masks, the model generates from scratch using full-image masking while allowing flexible region control to reflect user intent.

We also construct a diverse dataset without requiring segmentation or pose annotations and employ prompts from a large multimodal model to guide garment fit and style. Experimental results demonstrate that UP-VTON outperforms existing methods in flexibility, controllability, and visual realism, enabling high-fidelity and modality-free try-on synthesis.

1. Introduction

The primary objective of image-based virtual try-on (VTON) is to seamlessly and realistically overlay a given clothing image onto a target person image such that the resulting image retains both the visual features of the person and the specific details of the garment. Traditional VTON approaches [3, 7, 11, 12, 16, 31, 36, 38] typically involve two main stages: transforming the clothing image to align with the target's pose and body shape, and synthesizing the transformed clothing onto the person image. These methods rely on paired datasets consisting of garment and corresponding person images. A warping network learns the semantic correspondence between the garment and the body, followed by a generator that integrates the warped garment with the person image.

Despite the progress made by these methods, several limitations remain. In particular, current VTON methods [3, 8, 16, 36] often struggle to synthesize realistic try-on images in scenarios involving complex backgrounds or dynamic poses, due to the difficulty of collecting matched datasets across diverse environments. Moreover, traditional approaches [19, 40] frequently produce unrealistic outputs such as misaligned garments or visual artifacts, especially when the garment styles or poses deviate significantly from those in the training data. To overcome these challenges, recent works [4, 5, 8, 13, 22, 32, 33, 37, 43] have leveraged large-scale pre-trained diffusion models [27], which are known for their powerful capabilities in generating high-fidelity and high-resolution images. Originally developed for realistic human image synthesis, diffusion models have shown promising potential in virtual try-on tasks. Some diffusion-based methods [4, 13, 33, 37] incorporate architectures such as ReferenceNet and utilize denoising Unets combined with attention mechanisms to enhance the interaction between garment and person features. These approaches offer more robust and adaptive solutions that overcome the limitations of traditional pipelines.

Additionally, models such as Any2AnyTryon [9] and CatVTON [5] have further extended the capabilities of diffusion-based VTON. Any2AnyTryon supports flexible, prompt-driven virtual try-on synthesis based on text instructions and model garment images, even without masks, poses, or segmentation. CatVTON proposes a simpler mechanism that removes prompt inputs from the virtual try-on pipeline and utilizes a straightforward spatial concatenation strategy to combine garment and human features, achieving high-quality results with a more minimal structure.

However, existing methods fail to flexibly integrate both the presence and absence of masks and prompts, often relying on additional inputs such as keypoints, or requiring separately trained models for different input configurations—thereby limiting their flexibility and scalability. To overcome this limitation, we propose UP-VTON, a unified framework trained on a custom-built dataset. UP-VTON enables the synthesis of high-quality try-on results between a reference garment image and a target person image, while supporting a wide range of control modes—including both mask-based and text-guided editing.

Our approach leverages the multimodal attention mechanism of the Diffusion Transformer (DiT) [24] to jointly model semantic relationships among text descriptions, binary masks, and image patches. This design allows for flexible and effective conditioning in image editing tasks, regardless of whether the input guidance is textual or spatial.

In addition, we introduce an in-context editing mechanism that interprets the reference garment not as a static input but as a form of contextual content. This enables meaningful interaction between the garment and its surrounding visual scene, allowing the model to capture inherent correlations implicitly. Consequently, UP-VTON preserves the identity of the target person and seamlessly blends the in-

serted garment into the output image.

To implement this, we design a prompting strategy called Triptych Prompting, inspired by the classical art format consisting of three panels. In our method, the left panel contains the reference garment image, the center panel presents the target model image, and the right panel—defined by a binary mask—is the region to be completed via inpainting. A text prompt is provided alongside this triptych, describing the intended styling and scene context.

Finally, the right panel is generated through a text- and image-conditioned inpainting process, guided by pre-trained encoders. This process enables UP-VTON to generate coherent and realistic try-on results that reflect both the garment's visual identity and the user-defined context.

We evaluate our proposed model on two benchmark datasets: VITON-HD [3] and upper body in DressCode [21]. Experimental results consistently show that our method achieves state-of-the-art performance across a wide range of virtual try-on scenarios. Notably, as illustrated in Figure 1, the model also performs well on in-the-wild images, despite being trained exclusively on benchmark datasets. In summary, our main contributions are as follows.

- We propose UP-VTON, a unified DiT-based architecture for virtual try-on that operates on the target and garment images alone, while supporting both mask-agnostic control and prompt-based conditioning.
- To preserve garment fidelity and subject identity, we introduce Triptych Prompting, an in-context editing strategy that integrates reference garments into target images in a context-aware manner.

2. Related Work

2.1. Diffusion Transformer

Diffusion Transformer (DiT) [24] is a recently proposed architecture that incorporates transformer-based designs into diffusion models for image generation tasks. While conventional diffusion models typically rely on U-Net backbones [28], DiT replaces the convolutional hierarchy with a fully transformer-based structure, inspired by the success of Vision Transformers (ViT) [6] in image understanding.

Unlike convolutional backbones that primarily capture local spatial patterns, the self-attention mechanism in DiT enables modeling of global contextual dependencies across the entire image. This makes DiT particularly effective in tasks requiring long-range reasoning, such as layout consistency, cross-object relationships, and multi-modal alignment (e.g., between text and image).

Within diffusion models, DiT operates by replacing the denoising U-Net with a transformer module that takes as input a sequence of image tokens and time-step embeddings. Conditioning signals—such as class labels, text embeddings, or image guidance—can be incorporated through cross-attention layers. Recent studies have shown that DiT outperforms U-Net-based diffusion models on class-conditional image generation benchmarks such as ImageNet, while also providing greater architectural flexibility.

Owing to its strong capacity for cross-modal attention and semantic alignment, DiT has been recently applied to multimodal generation tasks including text-to-image synthesis, image editing, and object insertion.

2.2. Virtual Try-On

Image-based virtual try-on (VTON) aims to synthesize realistic images of a target person wearing a given garment, while preserving the person's identity and maintaining visual coherence. Traditional GAN-based methods [7, 11, 31, 36] typically adopt a two-stage pipeline: first, a warping module deforms the garment to fit the target's pose and body shape; then, a GAN-based generator merges the warped garment into the person's image. For instance, CP-VTON [31] separates the warping and generation processes to enable stage-specific optimization, while PF-AFN [7] employs knowledge distillation and appearance flow constraints to enhance synthesis quality. GP-VTON [36] further improves garment-person consistency by combining local flow-based warping with global human parsing. Despite their effectiveness, these approaches depend heavily on warping accuracy and often struggle to generalize to diverse poses and complex backgrounds.

Recently, diffusion models have emerged as powerful alternatives for high-fidelity conditional image synthesis, offering improved generalization and detail preservation over GAN-based approaches. TryOnDiffusion [43] uses dual U-Net backbones to simultaneously process garment, person, and pose features, producing high-quality results at the cost of higher computation. Hybrid methods such as DCI-VTON [8] and LaDI-VTON [22] combine diffusion with GAN components to balance quality and efficiency. DCI-VTON conditions a pre-trained diffusion model on garment features via a warping network, whereas LaDI-VTON uses CLIP-based garment embeddings to guide latent diffusion.

Efforts to adapt diffusion models specifically for VTON have yielded several innovations. StableVITON [13] introduces a zero cross-attention block to model semantic alignment between garment and person features. MMTryon [42] extends flexibility through multimodal conditioning and multi-reference garment inputs. OOTDiffusion [37] encodes garment labels using CLIP [25] to enable fine-grained control over different body regions. IDM-VTON [4] incorporates a ReferenceNet and image encoder to enhance garment-person alignment.

A notable advancement in recent VTON research is in-context editing, where the reference garment is treated as part of the semantic and spatial context of the tar-

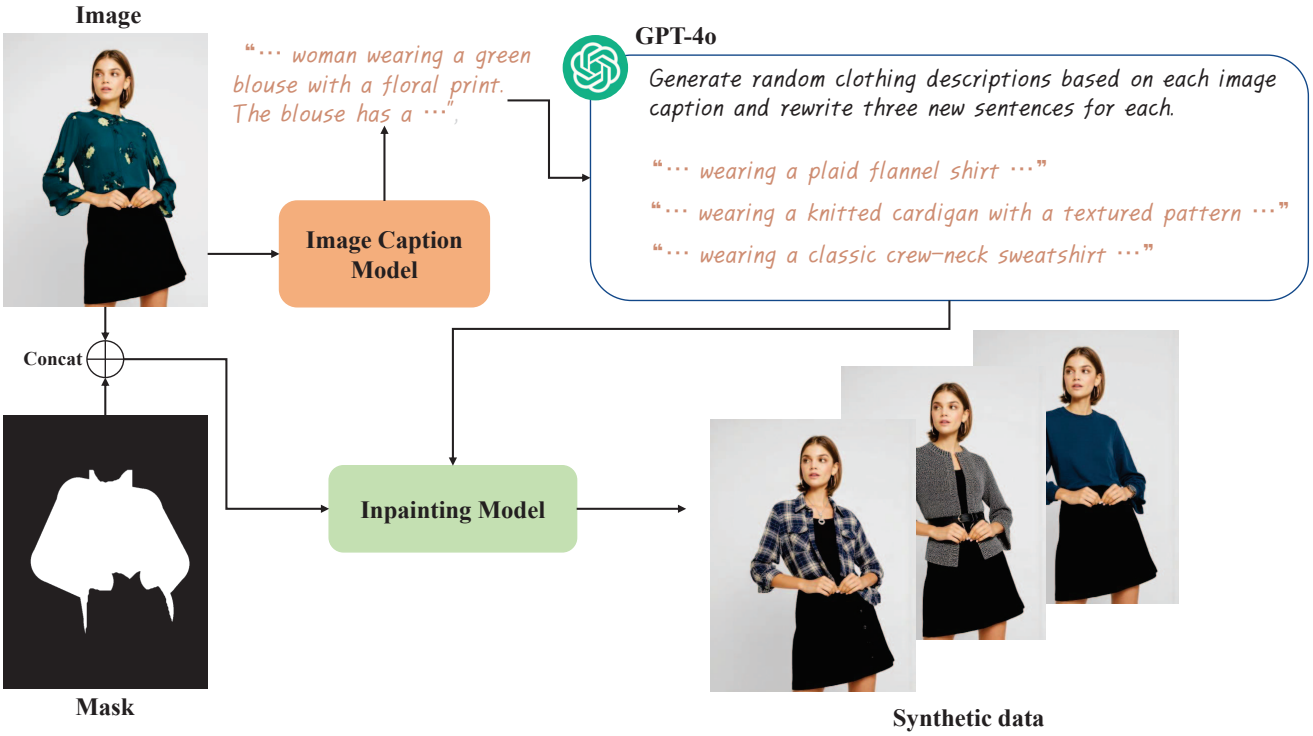

Figure 2. **Overview of the synthetic data generation pipeline**. Given an input person image, Florence2 [35] generates an initial image caption. The garment-related components of the caption are then rewritten or diversified using GPT-4o [1] to produce three semantically varied prompts. These text prompts are subsequently used to guide inpainting in the synthetic dataset construction process.

get image rather than as a separate entity. Methods like Any2AnyTryon [9] and CatVTON [5] adopt this paradigm by enabling implicit garment-person interaction during generation. However, both have limitations.

Any2AnyTryon, while offering flexibility through mask-free synthesis and text prompts, conditions generation on the entire image without explicit spatial segmentation. As a result, it often fails to preserve fine-grained identity cues such as facial structure, hairstyle, and pose alignment, limiting its use in scenarios where identity preservation is crucial.

On the other hand, CatVTON improves garment-person alignment via attention mechanisms but requires separately trained models for masked and mask-free settings, lacking a unified framework. Additionally, it does not support prompt-based control, preventing users from editing fit or style via textual instructions (e.g., "loose fit", "crop top"), thereby limiting interactivity and flexibility.

In contrast, our proposed framework, UP-VTON, integrates these capabilities by supporting both masked and mask-free input conditions, while also enabling semantic garment editing via textual prompts. Built on a single DiT-based architecture, UP-VTON employs triptych prompting and in-context learning to achieve identity-preserving, high-fidelity virtual try-on across diverse input configurations.

3. Method

3.1. Preliminary

FLUX.1 [15] is a state-of-the-art text-to-image generation model that combines the Flow Matching framework [18] with the Diffusion Transformer (DiT) architecture [24]. It has gained significant attention due to its strong capabilities in prompt interpretation, natural language grounding, and high-fidelity image synthesis.

FLUX.1 employs Rotary Positional Embedding (RoPE) [30] to capture relative positional information within latent representations. This enables the model to effectively learn spatial relationships and is essential for high-resolution image generation.

The Flow Matching component aligns the probabilistic flow from the noise distribution to the data distribution by learning a time-evolving velocity field. The training objective, known as conditional flow matching loss, is defined as:

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{t,p_t(z|\epsilon),p(\epsilon)}\left[\left|v_\Theta(z,t) - u_t(z|\epsilon)\right|^2\right]. \quad (1)$$

where $v_\Theta(z,t)$ represents the velocity field parameterized by the neural network, while $u_t(z|\epsilon)$ denotes the target conditional vector field derived from the noise ϵ. The expectation is taken over the diffusion timestep t, noisy sample

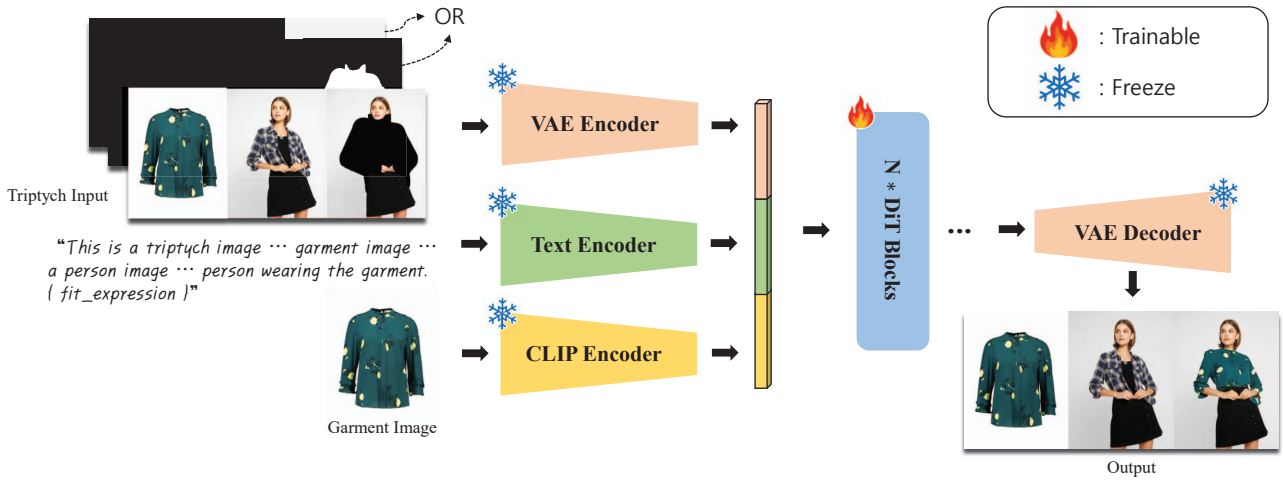

Figure 3. **Overview of the UP-VTON model framework**. The unified framework processes a triptych input—comprising a reference garment image, a source person image, and a corresponding mask—through a frozen VAE encoder. Semantic guidance is extracted from the garment image and text prompt via image and text encoders, then fused and passed to trainable DiT transformer blocks. The mask input can either represent the garment region or the entire image, enabling in-context learning for accurate and flexible virtual try-on with both mask-based and text-based prompts.

z, and base noise ϵ, ensuring that the model learns globally consistent generative trajectories across different noise conditions. Our method builds upon FLUX.1-Fill-Dev, an development version of inpainting-specialized variant derived from the FLUX.1 model.

3.2. Synthetic Dataset

We adopt a unified architecture that flexibly supports garment-region masks during virtual try-on. To ensure robustness in mask-free scenarios, we augment the training data with a synthetic dataset specifically designed for unmasked inputs.

To construct this dataset, we use DensePose [10] and SCHP [17] to generate garment segmentation masks from dressed person images. These masks are then used in an inpainting pipeline powered by FLUX.1-Fill-Dev, which synthesizes new images by removing and reconstructing clothing regions.

To guide the inpainting, we generate textual prompts via a two-step process (as illustrated in Figure 2): (1) Florence2 [35] is used to generate image captions from the original images. (2) The garment-related segments are then rewritten or diversified using GPT-4o [1] to create three distinct caption variants per image. These are used as prompts during inpainting.

This augmentation increases the dataset size fourfold and prepares the model to handle mask-free conditions effectively.

To support controllable garment fitting, we extract "fit" descriptions from the augmented captions. Using GPT-4o, we isolate and simplify these descriptions into concise expressions, denoted as $\{fit_expression\}$. Each $\{fit_expression\}$ is embedded into a templated prompt of the form:

"This is a triptych image composed of three connected images: the left panel shows a garment image, the center shows a person image, and the right panel shows the person wearing the garment image, $\{fit_expression\}$."

3.3. UP-VTON

As illustrated in Figure 3, our UP-VTON model performs virtual try-on using three inputs: a reference garment image, a target person image, and a control prompt (which includes a mask and a text instruction). The goal is to generate an output image in which the garment is seamlessly transferred to the target person, while preserving the identity and satisfying the prompt conditions.

The model employs a triptych-based in-context editing framework, arranging the inputs into three panels:

$$I_{\text{triptych}} = [\text{seg}(I_{\text{ref}}); I_{\text{syn}}; I_{\text{src}}^{\text{masked}}]. \tag{2}$$

Let $\text{seg}(\cdot)$ denote the garment segmentation process. To isolate the garment region, we apply Grounding-DINO [23] and SAM [14] to remove background elements from I_{ref}. Here, $[,;,]$ denotes the channel-wise concatenation operator. I_{ref} is the reference garment image, I_{syn} is a synthetic person image derived from the source image I_{src}, and $I_{\text{src}}^{\text{masked}}$ represents the masked person image used for inpainting. If no garment mask is provided, this third panel is substituted with an all-zero image.

We construct a binary mask:

$$M_{\text{triptych}} = [\mathbf{0}_{h\times w};\ \mathbf{0}_{h\times w};\ M_{h\times w}], \tag{3}$$

Figure 4. Qualitative comparison on a DressCode [21] dataset. These results are obtained without using any mask input in our model.

where $\mathbf{0}_{h\times w}$ corresponds to the reference and source panels, and M is the garment-region mask. If the mask is unavailable, M is set to all ones to allow free-form generation. This mechanism enables flexible control over the use of masks at both training and inference time.

Semantic features are extracted from the reference image using a CLIP image encoder and fused with text prompt embeddings obtained from a T5-based [26] text encoder for enhanced language understanding.:

$$\mathrm{C} = [E_I(I_{\text{ref}}); E_T(prompt)], \tag{4}$$

where E_I and E_T denote the image and text encoders, respectively.The resulting multimodal embedding C is injected into the text branch of the DiT architecture, enabling unified conditioning across visual and textual domains while preserving identity and enhancing controllability. The conditional input for training is structured as:

$$X = [x; concat[I_{triptych}, M_{triptych}]; C], \tag{5}$$

and the conditional flow matching loss becomes:

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{t,p_t(z|\epsilon),p(\epsilon)}\left[|v_\Theta(X,t) - u_t(z|\epsilon)|^2\right], \tag{6}$$

where $concat$ denotes channel-wise concatenation.

4. Experiment

Our method is based on FLUX.1-Fill-Dev, a publicly released development version of the DiT-based inpainting

Method	Extra Input	VITON-HD						DressCode (Upper Body)					
		Paired				Unpaired		Paired				Unpaired	
		FID↓	KID↓	SSIM↑	LPIPS↓	FID↓	KID↓	FID↓	KID↓	SSIM↑	LPIPS↓	FID↓	KID↓
DCI-VTON [8]	M,D,P	9.401	4.5474	0.8612	0.0605	12.5192	5.2497	-	-	-	-	-	-
LADI-VTON [22]	M,K	11.3607	7.254	0.8644	0.0732	14.6293	8.7418	-	-	-	-	-	-
StableVITON [13]	M,D	6.4071	0.9436	0.8563	0.0905	11.0347	3.9118	-	-	-	-	-	-
IDM-VTON [4]	M,D	5.7752	0.7319	0.8464	0.0603	9.8703	1.121	7.5249	1.1223	0.9298	0.0368	11.7389	1.6249
OOTDiffusion [37]	M	9.3211	4.0963	0.8207	0.0874	12.3838	4.6797	9.4124	1.0326	0.9078	0.0512	14.5650	2.5646
CatVTON [5]	M	5.4417	0.4306	0.8700	0.0575	9.0319	1.0907	9.4052	2.2195	0.9177	0.0559	13.6483	3.2863
Any2AnyTryOn [9]	-	6.9340	0.7387	0.8387	0.0877	8.9650	0.981	13.9546	6.0374	0.9092	0.0758	15.2549	6.8829
Ours (w/o M)	-	5.8801	0.7467	0.8414	0.0648	**7.9875**	**0.6912**	7.6663	1.1195	0.9098	0.0403	<u>12.8535</u>	<u>2.1122</u>
Ours (w/ M)	M	**5.2597**	**0.4206**	**0.8882**	**0.0564**	8.6822	0.7375	**7.4621**	**1.1195**	**0.9439**	**0.0349**	13.2017	2.1638

Table 1. Quantitative comparison on VITON-HD and DressCode (Upper Body) under both paired and unpaired settings. Our method shows competitive or superior performance without requiring extra inputs beyond image and prompt. In the "Extra Input" column, M denotes mask input. And D denotes dense pose map, P denotes parsing map, and K denotes keypoints of person. The highest-performing result is highlighted in **bold**, and the second-best is marked with an <u>underline</u> for clarity.

Setting	FID ↓	KID ↓	SSIM ↑	LPIPS ↓
w/o synthetic data	6.7501	0.9110	0.8132	0.0916
w/o triptych	6.3080	0.7088	0.7293	0.1458
full	**5.8801**	**0.7467**	**0.8414**	**0.0648**

Table 2. Ablation study on the proposed components.

model from FLUX.1 [15]. We use the SigLIP [39] image encoder and T5 [26] text encoder. All images are resized to a resolution of 832×624 (height × width) for training. Throughout all training stages, we adopt the Prodigy [20] optimizer with a weight decay of 0.01 for training transformer blocks.

For training data, we utilize the VITON-HD [3] dataset and the *Upper Body* subset of DressCode [21]. For both datasets, we generate three times the amount of synthetic data using the method described in Section 3.2, which is used to enhance the training corpus. When constructing the input triptych, we apply a 0.5 probability of using the clothing region mask. If the clothing region mask is not used, the entire third panel of the triptych is fully masked.

4.1. Qualitative Results

Figure 4 shows qualitative comparisons on the DressCode [21] dataset against four recent state-of-the-art methods. As seen in the figure, our model exhibits clear advantages in handling complex textures, producing noticeably fewer artifacts, better preservation of logos and patterns, and reduced loss of fine details. In addition, non-garment regions such as the background and hand shapes are rendered more faithfully, preserving the identity of the source image. As demonstrated in the first row, our method naturally synthesizes garments with intricate structures, such as lace hems. Furthermore, as shown in the fourth row, it successfully handles garments with uncommon and challenging shapes, generating high-quality try-on results.

4.2. Quantitative Results

For quantitative evaluation, we use SSIM [34] and LPIPS [41] in paired settings and FID [29] and KID [2] scores for both paired and unpaired settings to assess realism. Table 1 shows the results of UP-VTON and comparative methods on a single-dataset setup. UP-VTON outperforms all methods across unpaired settings ranked second or higher in paired. Additionally, IDM-VTON [4] and CatVTON [5] also exhibit strong performance. Notably, IDM-VTON demonstrates that the use of additional dense pose information can aid in compensating for unreasonable mask artifacts in certain input images.

4.3. Ablation study

To evaluate the effects of the proposed components, we conducted an ablation study using the VITON-HD [3] dataset. Table 2 presents quantitative results as each component was incrementally introduced. This experiment was conducted primarily using mask-free inputs. First, we observed a significant performance drop when training the model without augmented data. As discussed in Section 3.2, this result indicates that in the absence of a mask, the influence of the source model becomes overly dominant, hindering the effective incorporation of reference garments. Thus, incorporating augmented data helps mitigate this imbalance and enables the training of a more scalable and generalizable model. Second, we compared the performance without using the triptych input. In this case, the model used a diptych input, where the reference garment image was processed only through the image encoder. This approach failed to capture fine-grained details of the garment—such as logos or complex patterns—resulting in noticeable performance degradation.

5. Conclusion

The proposed UP-VTON framework represents a significant advancement in virtual try-on by introducing a unified, mask-agnostic solution capable of handling a wide range of scenarios. By leveraging innovative techniques such as synthetic dataset augmentation and a triptych input design, the method effectively enhances the quality of virtual garment generation. Extensive experiments demonstrate the effectiveness of UP-VTON, highlighting its superior flexibility in accommodating user inputs and its ability to control the presence or absence of mask guidance. The framework consistently generates high-fidelity and realistic try-on results across various conditions. Overall, UP-VTON offers strong flexibility, scalability, and generalization capability, marking an important step forward for both research and real-world applications in the field of virtual try-on.

Acknowledgments

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government (No.25ZB1100, Core Technology Research for Self-Improving Integrated Artificial Intelligence System, 80%) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2022-II220907, Development of AI Bots Collaboration Platform and Self-organizing AI, 20%).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 4, 5

[2] Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans, 2021. 7

[3] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *CVPR*, pages 14131–14140, 2021. 1, 2, 3, 7

[4] 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, 7

[5] 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, 2024. 2, 4, 7

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[7] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *CVPR*, pages 8485–8493, 2021. 2, 3

[8] Junhong Gou, Siyu Sun, Jianfu Zhang, Jianlou Si, Chen Qian, and Liqing Zhang. Taming the power of diffusion models for high-quality virtual try-on with appearance flow. *Proceedings of the ACM International Conference on Multimedia*, 2023. 2, 3, 7

[9] Hailong Guo, Bohan Zeng, Yiren Song, Wentao Zhang, Chuang Zhang, and Jiaming Liu. Any2anytryon: Leveraging adaptive position embeddings for versatile virtual clothing tasks, 2025. 2, 4, 7

[10] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, pages 7297–7306, 2018. 5

[11] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *CVPR*, pages 7543–7552, 2018. 2, 3

[12] Thibaut Issenhuth, Jérémie Mary, and Clément Calauzenes. Do not mask what you do not need to mask: a parser-free virtual try-on. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 619–635. Springer, 2020. 2

[13] Jeongho Kim, Gyojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. *arXiv preprint arxiv:2312.01725*, 2023. 2, 3, 7

[14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 5

[15] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 4, 7

[16] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *ECCV*, pages 204–219. Springer, 2022. 2

[17] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3260–3271, 2020. 5

[18] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 4

[19] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 2

[20] Konstantin Mishchenko and Aaron Defazio. Prodigy: An expeditiously adaptive parameter-free learner. *arXiv preprint arXiv:2306.06101*, 2023. 7

[21] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: High-resolution multi-category virtual try-on. In *ECCV*, pages 2231–2235, 2022. 1, 3, 6, 7

[22] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. *arXiv preprint arXiv:2305.13501*, 2023. 2, 3, 7

[23] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 5

[24] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4172–4182, 2023. 2, 3, 4

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[26] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 6, 7

[27] 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. 2

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 3

[29] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, 2020. Version 0.3.0. 7

[30] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 4

[31] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *ECCV*, pages 589–604, 2018. 2, 3

[32] 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. 2

[33] Rui Wang, Hailong Guo, Jiaming Liu, Huaxia Li, Haibo Zhao, Xu Tang, Yao Hu, Hao Tang, and Peipei Li. Stablegarment: Garment-centric generation via stable diffusion. *arXiv preprint arXiv:2403.10783*, 2024. 2

[34] 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

[35] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks (2023). *URL https://arxiv.org/abs/2311.06242*, 2023. 4, 5

[36] 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 *CVPR*, pages 23550–23559, 2023. 2, 3

[37] Yuhao Xu, Tao Gu, Weifeng Chen, and Chengcai Chen. Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. *arXiv preprint arXiv:2403.01779*, 2024. 2, 3, 7

[38] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *CVPR*, pages 7850–7859, 2020. 2

[39] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 7

[40] 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

[41] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 7

[42] 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, 2024. 3

[43] 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. 2, 3