

OmniVTON: Training-Free Universal Virtual Try-On

Zhaotong Yang¹, Yuhui Li¹, Shengfeng He², Xinzhe Li¹, Yangyang Xu³, Junyu Dong¹, Yong Du^{1*}
¹Ocean University of China,
²Singapore Management University, ³Harbin Institute of Technology (Shenzhen)



Figure 1. We propose OmniVTON, a training-free universal virtual try-on framework that unifies both in-shop and in-the-wild scenarios while preserving garment details and ensuring pose consistency.

Abstract

Image-based Virtual Try-On (VTON) techniques rely on either supervised in-shop approaches, which ensure high fidelity but struggle with cross-domain generalization, or unsupervised in-the-wild methods, which improve adaptability but remain constrained by data biases and limited universality. A unified, training-free solution that works across both scenarios remains an open challenge. We propose OmniVTON, the first training-free universal VTON framework that decouples garment and pose conditioning to achieve both texture fidelity and pose consistency across diverse settings. To preserve garment details, we introduce a garment prior generation mechanism that aligns clothing with the body, followed by continuous boundary stitching technique to achieve fine-grained texture retention. For precise pose alignment, we utilize DDIM inversion to capture structural

cues while suppressing texture interference, ensuring accurate body alignment independent of the original image textures. By disentangling garment and pose constraints, OmniVTON eliminates the bias inherent in diffusion models when handling multiple conditions simultaneously. Experimental results demonstrate that OmniVTON achieves superior performance across diverse datasets, garment types, and application scenarios. Notably, it is the first framework capable of multi-human VTON, enabling realistic garment transfer across multiple individuals in a single scene. Code is available at <https://github.com/Jerome-Young/OmniVTON>.

1. Introduction

Image-based virtual try-on (VTON) transforms online shopping by seamlessly integrating garment images with target human bodies to generate natural-looking results that

*Corresponding author (csyongdu@ouc.edu.cn).

conform to body poses while preserving texture consistency. It enhances the shopping experience by reducing uncertainty and minimizing return rates.

Existing VTON methods are designed for either in-shop or in-the-wild scenarios. Supervised approaches [8, 18, 26, 40, 42] dominate in-shop settings, achieving high-fidelity synthesis using paired training data but struggling with cross-domain/-scenario generalization. Conversely, in-the-wild methods [11] leverage unsupervised learning to improve adaptability across diverse input sources (*e.g.*, Shop-to-Street, Model-to-Model, *etc.*) but remain constrained by data distribution biases and limited universality. Both paradigms rely on dedicated models trained for specific conditions, making large-scale dataset construction across all garment categories, styles, and human poses highly impractical. This fragmentation underscores the need for a unified VTON framework that can generalize across domains without requiring additional training.

To enable a training-free VTON framework, two critical challenges must be addressed:

i) *Fine-grained Texture Consistency*: Without a dedicated training phase, it is difficult to establish garment-body alignment while preserving intricate texture details. Conventional methods rely on learned deformation priors, which are unavailable in a training-free setting.

ii) *Human Pose Alignment*: Existing methods condition on keypoints [5] or DensePose maps [19], requiring retraining for cross-modal feature fusion. Without explicit pose supervision, training-free approaches would struggle with pose consistency, especially for garments with ambiguous structures like sleeveless vests.

To tackle these issues, we propose *OmniVTON*, a training-free virtual try-on framework that leverages off-the-shelf diffusion models through a progressive garment adaptation mechanism. For texture preservation, we introduce Structured Garment Morphing (SGM), which ensures seamless garment-body integration while maintaining fine-grained texture details. First, a pseudo-person image is generated via semantic-guided garment completion. Then, multi-part semantic correspondence between the pseudo-person and the source person is established using a predicted segmentation map and skeleton data. Finally, localized transformations dynamically adjust different garment regions to achieve an anatomically accurate alignment, producing a structurally coherent adaptation result. To address inconsistencies along garment boundaries, we propose Continuous Boundary Stitching (CBS), which refines the transitions between segmented regions to ensure seamless integration. By leveraging semantic interactions between the latent features of the original garment image and the garment-infused image, CBS eliminates harsh edges and discontinuities, preserving the visual realism of the final synthesis.

For pose information injection, a naive approach is to

directly apply DDIM Inversion [34], which preserves structural information by replacing the initial random noise with inversion noise from the source person. However, this also introduces unwanted texture contamination. To address this, we propose Spectral Pose Injection (SPI), which selectively integrates pose cues while suppressing texture interference. By leveraging spectral analysis in the latent space, SPI retains low-frequency inversion noise for structural consistency while replacing high-frequency components with random noise to enhance generative flexibility. This frequency-aware modulation maintains pose fidelity while preventing residual texture artifacts.

Comprehensive experiments demonstrate that *OmniVTON* surpasses existing methods across multiple benchmarks, both qualitatively and quantitatively, producing high-fidelity try-on results while offering new insights into virtual try-on. Additionally, it showcases strong generalizability across various scenarios, datasets, and clothing types. The key contributions of this work include:

- We propose *OmniVTON*, a training-free universal VTON framework that unifies in-shop and in-the-wild scenarios, significantly advancing the state of the art.
- We introduce Structured Garment Morphing, ensuring fine-grained texture preservation and seamless garment adaptation across diverse clothing types and scenarios.
- We develop Spectral Pose Injection and Continuous Boundary Stitching to effectively integrate pose information and refine textures, producing pose-consistent and texture-coherent try-on results.
- Our method achieves state-of-the-art results across multiple evaluation metrics, demonstrating superior quality, generalizability, and scalability. Notably, it is the first to enable multi-human VTON, facilitating realistic garment transfer across multiple individuals.

2. Related work

Garment Warping. Garment warping is a core component of virtual try-on systems, as it ensures accurate body alignment and texture preservation. Early approaches [8, 20] relied on Thin Plate Spline [4] (TPS) deformation with sparse control points, but their low-dimensional parametric representation failed to capture complex pose variations. Subsequent works [21, 23, 39, 42] introduced dense optical flow [45] to enable pixel-level semantic alignment, achieving improved deformation quality, yet still requiring paired training data. To alleviate this constraint, *PastaGAN++* [38] proposed patch-routed disentanglement for unpaired training. However, most existing methods are tailored to specific scenarios, limiting their adaptability across diverse input sources.

While out-of-distribution generalization [33] and domain-agnostic modeling [13] have attracted increasing interest in other areas, these capabilities remain underex-

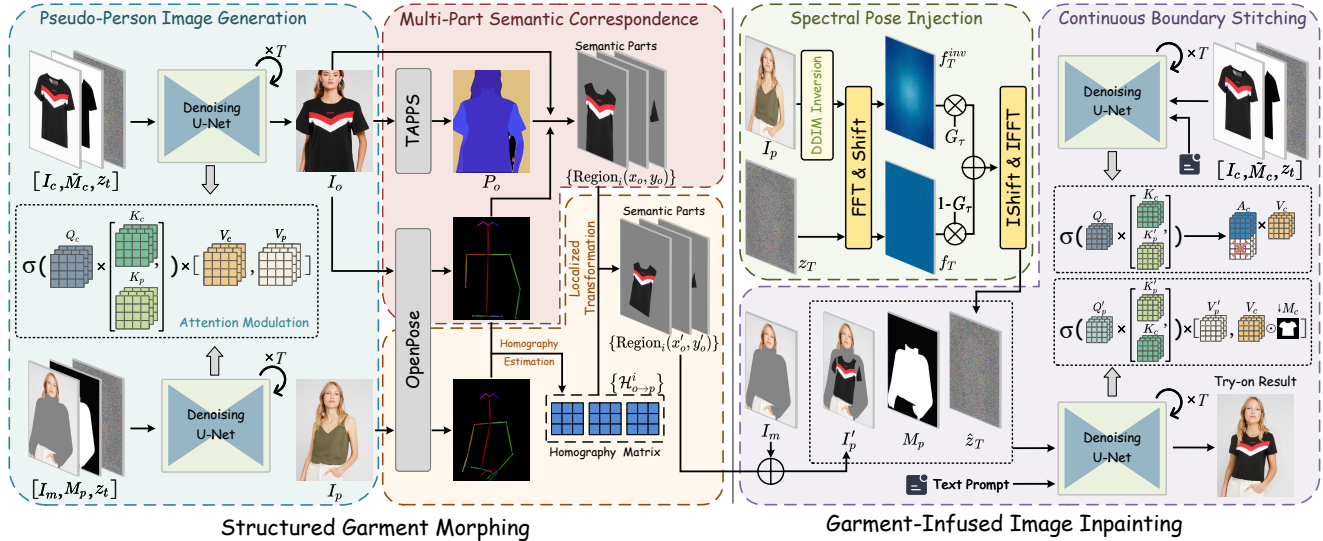


Figure 2. Overview of OmniVTON. It consists of two main steps: 1) Utilize pseudo-person image I_o to achieve multi-part deformation, generating adapted clothing. 2) Integrate this prior with clothing-agnostic image I_m to create garment-infused image I'_p , which is concatenated with pose-encoded noise \hat{z}_T as input, thereby obtaining refined try-on result through the Continuous Boundary Stitching mechanism.

plored in garment warping. OmniVTON addresses this gap by enabling training-free and universal garment adaptation through skeletal guidance. Compared to StreetTryOn [11], which supports cross-domain try-on, our method avoids dense warping’s failure modes in in-shop lower-garment cases and better preserves garment structure. As a result, OmniVTON achieves superior versatility and fidelity across a wide range of virtual try-on tasks.

Image-Based Virtual Try-On. Implicit warping-based virtual try-on methods [6, 9, 25, 30, 40, 43] have recently gained attention for their ability to jointly model garment deformation and human-body synthesis using diffusion models’ powerful semantic correspondence capabilities. Ladi-VTON [30], for instance, employs textual inversion [14] to map garment textures to text-based conditions, but its reliance on textual ambiguity results in insufficient control over garment details. To improve garment-body interactions, IDM-VTON [9] and StableVTON [25] incorporate advanced attention mechanisms, yet the absence of explicit deformation constraints often leads to geometric misalignment and texture inconsistencies, particularly in open-domain scenarios. Unlike these methods, OmniVTON provides direct texture guidance through structured garment priors during the inpainting stage, ensuring precise alignment and preserving fine-grained garment details. Its training-free paradigm enables universal applicability across diverse datasets, scenarios, and garment categories.

Exemplar-Based Image Inpainting. Both exemplar-based image inpainting [7, 27, 41] and virtual try-on require accurate feature transfer from reference images to target regions. PBE [41] trains an image encoder to align visual and textual semantics, while AnyDoor [7] enhances texture represen-

tation by injecting multi-level high-frequency features into U-Net [32]. However, excessive high-frequency retention often causes style inconsistencies in the generated outputs. In contrast, text-driven inpainting methods [24, 46] suffer from limited information granularity, leading to texture distortion, whereas personalized approaches [3, 14] generate more identity-preserving text embeddings but require fine-tuning. Our proposed Spectral Pose Injection (SPI) offers a novel alternative by abandoning strict high-frequency constraints while leveraging OmniVTON to maintain garment identity consistency. By integrating structured pose-aware noise modulation, SPI ensures that try-on results conform precisely to the target person’s pose without sacrificing texture fidelity.

3. Approach

Given a garment-contained¹ image I_c and a target person image I_p , our goal is to seamlessly transfer the indicated garment onto the corresponding semantic region of I_p without any training. To this end, we tackle two key challenges:

- Warping the given garment in a training-free manner.
- Preserving the original person’s pose while inpainting the cloth-agnostic image, also without training.

As illustrated in Fig. 2, OmniVTON follows a two-step workflow. First, it morphs the target garment to create a garment prior aligned with the human body. Then, using this prior and pose-encoded noise, it progressively refines the boundary of the garment and completes the garment-infused image, ensuring a coherent and pose-matching result.

¹A garment-contained image I_c refers to either a standalone garment or a person wearing it in diverse backgrounds.

3.1. Structured Garment Morphing

We propose Structured Garment Morphing (SGM) to accurately deform the target garment. Unlike TPS and Flow-based methods [8, 39], which require retraining for different domains, SGM leverages skeletal information and parsing maps to constrain garment morphing. It establishes a one-to-one mapping between the target garment and the original worn person images using correspondences between the target and source person. While this approach is naturally applicable in Non-Shop-to-X settings, the Shop-to-X setting faces challenges when only the garment image is available, leading to failures in keypoint detection and parsing due to the lack of parseable human body structure. To ensure universality, SGM’s first task is to generate a pseudo-person image from the garment to extract reliable information.

Pseudo-Person Image Generation. We empirically found that text-driven image generation paradigm often fails to produce the desired pseudo-human image I_o , likely due to weak controllability or the need for carefully crafted prompts. Instead, we propose modulating attention outputs for generation. Specifically, we first relocate the target garment to the semantically relevant region of the source person’s body, establishing an initial spatial correspondence. To inject the person’s semantic features, we parallel-denoise both the garment-conditioned noise z_t , concatenated with the garment image I_c and the inverted cloth mask \tilde{M}_c , which indicates the region to be generated. We also apply the same noise conditioned on the worn person’s information, concatenated with the cloth-agnostic image I_m and the agnostic mask M_p , integrating the latter’s key and value into the self-attention layers of the former:

$$f_c = \text{Softmax} \left(\frac{Q_c \cdot [K_c \parallel K_p]^\top}{\sqrt{d}} \right) [V_c \parallel V_p], \quad (1)$$

where K_p and V_p are the key and value matrices of the worn person image, Q_c , K_c , and V_c represent the query, key, and value of the garment image, and \parallel denotes tensor concatenation along the spatial dimension. Since the key and value encode an image’s spatial layout and content information [35], this mechanism effectively incorporates human body semantics into the pseudo-person generation process while preserving the target garment’s texture.

Multi-Part Semantic Correspondence. After obtaining the target person image, we use skeleton and parsing to establish multi-part semantic correspondence between the target garment and the original worn person image. First, we define a set of N human semantic regions and use OpenPose [5] for semantic disentanglement on both images.

Taking an upper garment as an example, we define five semantic regions: torso, left and right upper arms, and left and right lower arms. Using the 25 keypoints predicted by OpenPose, we construct bounding boxes $\{B_o^i\}_{i=1}^5$ to en-

compass all keypoints of each region in I_o . Likewise, we obtain the corresponding bounding boxes $\{B_p^i\}_{i=1}^5$ for I_p .

To avoid interference from overlapping parts, we then apply a human part segmentation map P_o , generated by TAPPS [12], to isolate pixels corresponding to each region:

$$\mathbb{I}_{\text{Region}_i}(x_o, y_o) = \begin{cases} 1, & \text{if } (x_o, y_o) \in P_o^i \cap B_o^i, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

where $\mathbb{I}(\cdot)$ indicates the indicator function, and (x_o, y_o) represents the pixel coordinates of I_o . Note that although the generated pseudo-person image may not always capture the entire human body, our relocate operation ensures that at least the outer body, including the garment, is covered. This is sufficient to establish the multi-part correspondence needed for the subsequent localized transformation.

Localized Transformations. For the corner points of each bounding box pair $\{B_o^i, B_p^i\}$, we optimize the homography matrix $\mathcal{H}_{o \rightarrow p}^i \in \mathbb{R}^{3 \times 3}$ using the Levenberg-Marquardt algorithm [15]. Then, a piecewise perspective transformation is applied to I_o to align it with the source human geometry:

$$\begin{bmatrix} x'_o \\ y'_o \\ 1 \end{bmatrix} = \sum_{i=1}^5 \mathbb{I}_{\text{Region}_i}(x_o, y_o) \mathcal{H}_{o \rightarrow p}^i \begin{bmatrix} x_o \\ y_o \\ 1 \end{bmatrix}, \quad (3)$$

where (x'_o, y'_o) represents the pixel coordinates at (x_o, y_o) after morphing. In this way, we obtain a coarse deformed garment I_w , which, through multi-region stitching, serves as an effective prior for the subsequent step of garment-infused image inpainting.

The one-to-one mapping characteristic of SGM eliminates the need for training; however, this multi-region stitching approach results in discontinuities along the boundaries of the morphed garment. We will discuss methods for boundary refinement in Sec. 3.3.

3.2. Spectral Pose Injection

The VTON task requires high human pose fidelity, as relying solely on the local deformation from garment morphing is insufficient for full-body pose alignment. This issue is especially noticeable with garments of ambiguous structures, like sleeveless vests or shorts. While skeleton-based conditioning in diffusion models can improve pose controllability, combining it with other conditions, such as text prompts, may cause the model to overfit certain conditions while neglecting others [22]. Alternatively, we apply DDIM Inversion [34] to reverse-map I_p into latent space, obtaining noise z_T^{inv} that preserves source human body structure. However, z_T^{inv} also retains the source garment’s texture, which may conflict with the texture generation of the target garment during image inpainting.

To address this, we propose Spectral Pose Injection (SPI), inspired by our spectral analysis in latent space. As

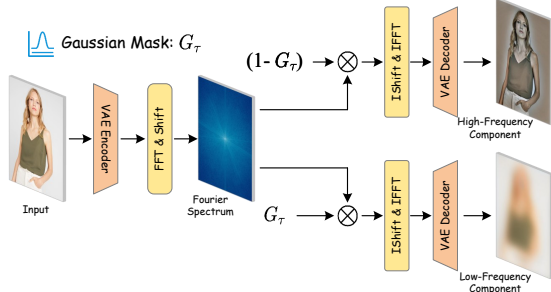


Figure 3. Visualization of distinct spectral bands in latent space.

shown in Fig. 3, the human latent, when decomposed via the Fast Fourier Transform (FFT) into low- and high-frequency components, exhibits distinct characteristics in reconstructing the original image. The low-frequency component captures only the coarse human silhouette, which adequately preserves pose information, while the high-frequency component retains both pose and fine garment textures.

The core idea of SPI is to retain the low-frequency structural information from the inverted noise while leveraging the high-frequency components of random noise to enhance generative flexibility. Specifically, we first apply FFT and centralization to both z_T^{inv} and a random noise z_T :

$$f_T^{inv} = \text{Shift}(\mathcal{F}(z_T^{inv})), \quad f_T = \text{Shift}(\mathcal{F}(z_T)), \quad (4)$$

where $\mathcal{F}(\cdot)$ denotes the Fourier transform, and $\text{Shift}(\cdot)$ shifts the low-frequency components to the center, facilitating spectral decoupling.

Next, we perform frequency-domain weighted fusion using a Gaussian low-pass mask G_τ , where τ controls the cut-off frequency:

$$\hat{f}_T = G_\tau \odot f_T^{inv} + (1 - G_\tau) \odot f_T, \quad (5)$$

where \odot denotes element-wise multiplication. The mask G_τ ensures that the low-frequency pose information from z_T^{inv} is preserved while injecting the high-frequency randomness from z_T to eliminate texture residuals.

Finally, we apply the inverse Fourier transform to the fused spectrum \hat{f}_T to obtain the mixed initial noise:

$$\hat{z}_T = \mathcal{F}^{-1}(\text{Shift}^{-1}(\hat{f}_T)). \quad (6)$$

3.3. Continuous Boundary Stitching

During the inpainting stage, the diffusion model takes the concatenation of mixed noise \hat{z}_T , the garment-infused image I'_p , and the cloth-agnostic mask as input to generate the final try-on image. Since I_w is assembled by morphing and stitching garment regions, its boundaries may exhibit texture discontinuities. These artifacts can be misinterpreted by the diffusion model as inherent garment details, resulting in unrealistic seams or misaligned patterns in the output.

To address this, we propose the Continuous Boundary Stitching (CBS) mechanism, which leverages bidirectional

semantic context information between I_c and I'_p to improve boundary continuity during the inpainting process. Similar to attention modulation in pseudo-person image generation, CBS operates by manipulating the self-attention outputs. The key difference is that CBS enables dual-path feature exchange, where the interaction from the I_c -path to the I'_p -path is defined as follows:

$$f'_p = \text{Softmax} \left(\frac{Q'_p \cdot [K'_p \parallel K_c]^\top}{\sqrt{d}} \right) [V'_p \parallel (V_c \cdot \downarrow M_c)], \quad (7)$$

where \downarrow represents downsampling M_c to match the dimension of V_c , aiming to suppress interference from the background information of I_c . The operation in Eq. (7) allows the query Q'_p to match the target garment texture, thereby bridging discontinuities caused by multi-region stitching.

In addition, we also adjust the self-attention output of I_c by using the key from the I'_p -path:

$$A_c = \text{Softmax} \left(\frac{Q_c \cdot [K_c \parallel K'_p]^\top}{\sqrt{d}} \right), \quad f_c = A_c[:, 1:n] \cdot V_c, \quad (8)$$

where $A_c \in \mathbb{R}^{n \times 2n}$ denotes the self-attention map. This operation enhances the similarity between the attention maps of I_c and I'_p , while suppressing dissimilar values. As a result, A_c retains its continuous boundary and adjusts to align with the layout of I'_p . This optimization further improves the information flow from the I_c -path to the I'_p -path, aiding boundary refinement in the subsequent time step. Note that we exclude V'_p to prevent texture interference from its discontinuities, so only the first n columns of the attention map are used in the output calculation.

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate OmniVTON on two in-shop datasets (VITON-HD [8], DressCode [29]) and one in-the-wild dataset (DeepFashion2 [16]). VITON-HD provides 2,032 upper garment-model test pairs, while DressCode spans three subcategories (upper, lower, and dresses) with a total of 5,400 test samples. For DeepFashion2, following the StreetTryOn benchmark [11], we constructed a 2,089-image test set covering four try-on scenarios: Shop-to-Street, Model-to-Model, Model-to-Street, and Street-to-Street. Input resolution was dynamically adjusted based on the target person source: 512×384 for VITON-HD/DressCode and 512×320 for DeepFashion2.

Baselines and Metrics. We compare OmniVTON against two baseline categories: exemplar-based image editing methods (PBE [41], AnyDoor [7], TIGIC [27], Cross-Image [1]) and traditional virtual try-on models (PWS [2], PastaGAN++ [38], GP-VTON [39], CAT-DM [43], D⁴-VTON [42], IDM-VTON [9], and StreetTryOn [11]).

Method	Year	FID _u ↓	FID _p ↓	SSIM _p ↑	LPIPS _p ↓
PBE [41]	2023 (CVPR)	19.230	17.649	0.784	0.227
AnyDoor [7]	2024 (CVPR)	14.830	9.922	0.796	0.164
TIGIC [27]	2024 (ECCV)	90.338	88.900	0.613	0.422
Cross-Image [1]	2024 (SIGGRAPH)	62.614	57.286	0.760	0.256
GP-VTON [39]	2023 (CVPR)	51.566	49.196	0.810	0.249
CAT-DM [43]	2024 (CVPR)	28.869	26.339	0.775	0.229
D ⁴ -VTON [42]	2024 (ECCV)	25.299	23.914	0.790	0.250
IDM-VTON [9]	2024 (ECCV)	23.035	20.460	0.812	0.147
Ours	-	9.621	7.758	0.832	0.145

Table 1. Quantitative comparisons on the VITON-HD dataset [8], where the subscript u and p indicates the unpaired and paired settings, respectively.

Among image editing methods, PBE and AnyDoor leverage large-scale pretraining for image inpainting, while TIGIC and Cross-Image utilize cross-image attention for localized editing. Traditional VTON models, except StreetTryOn, are scenario-specific: PWS and PastaGAN++ are trained on Model-to-Model datasets (DeepFashion [28] and UPT [37], respectively), whereas GP-VTON, CAT-DM, and others focus on Shop-to-Model settings.

Evaluation follows standard protocols, using Fréchet Inception Distance (FID) [31] to measure the similarity between generated try-on results and real image distributions. For VITON-HD and DressCode, which contain ground-truth images, we also employ Structural Similarity (SSIM) [36] and Learned Perceptual Image Patch Similarity (LPIPS) [44] to assess structural integrity and texture fidelity.

4.2. Comparison with State-of-the-Art Methods

Quantitative Evaluation. Tab. 1 presents the quantitative evaluation of OmniVTON on the VITON-HD dataset. To assess *cross-dataset* generalization, all VTON methods were tested using official checkpoints pre-trained on DressCode. OmniVTON outperforms the best-performing baseline by 0.020 in SSIM and 0.002 in LPIPS, confirming its superiority in pose preservation and appearance fidelity. More notably, our method reduces the FID metric by 5.209 in the unpaired setting, demonstrating exceptional cross-domain adaptability. Notably, while the second-best performer, AnyDoor, benefits from training on a dataset that includes VITON-HD samples, leading to favorable FID_u and FID_p scores, its structural accuracy remains constrained by the absence of geometric garment guidance.

To evaluate *cross-type* adaptability, we tested VTON methods using VITON-HD pre-trained models (which contain only upper garments) on the DressCode dataset, which includes diverse clothing types. As shown in Tab. 2, OmniVTON achieves substantial improvements across all metrics, outperforming both exemplar-based editing and VTON baselines with at least a 33.4% relative enhancement in

Method	Year	FID _u ↓	FID _p ↓	SSIM _p ↑	LPIPS _p ↓
PBE [41]	2023 (CVPR)	14.851	13.677	0.846	0.155
AnyDoor [7]	2024 (CVPR)	14.562	14.411	0.798	0.202
TIGIC [27]	2024 (ECCV)	64.117	63.531	0.749	0.319
Cross-Image [1]	2024 (SIGGRAPH)	38.438	34.917	0.841	0.161
GP-VTON [39]	2023 (CVPR)	44.753	44.469	0.843	0.218
CAT-DM [43]	2024 (CVPR)	13.678	12.028	0.858	0.125
D ⁴ -VTON [42]	2024 (ECCV)	22.390	21.435	0.841	0.152
IDM-VTON [9]	2024 (ECCV)	9.685	8.377	0.842	0.138
Ours	-	6.450	5.335	0.865	0.119

Table 2. Quantitative comparisons on the DressCode dataset [29], where the subscript u and p indicates the unpaired and paired settings, respectively.

	Shop-to-Street	Model-to-Model	Model-to-Street	Street-to-Street
	FID↓	FID↓	FID↓	FID↓
PBE [41]	81.538	20.181	62.664	36.556
AnyDoor [7]	50.893	24.235	51.861	35.139
TIGIC [27]	100.177	114.151	130.836	121.520
Cross-Image [1]	69.444	52.310	66.755	57.753
CAT-DM [43]	37.484	-	-	-
D ⁴ -VTON [42]	35.003	-	-	-
IDM-VTON [9]	42.282	-	-	-
PWS [2]	-	34.858	77.274	84.990
PastaGAN++ [38]	-	13.848	71.090	67.016
StreetTryOn [11]	34.054	12.185	34.191	33.039
Ours	33.919	8.983	33.450	23.470

Table 3. Quantitative comparisons on the StreetTryOn benchmark [11]. Virtual try-on methods use publicly available models trained on VITON-HD [8], while PWS [2] and PastaGAN++ [38] are trained on DeepFashion [28] and UPT [37], respectively. StreetTryOn results are taken from its original paper.

FID_u. This performance gain stems from the synergistic effects of Structured Garment Morphing (SGM) and Continuous Boundary Stitching (CBS), which collectively enhance robustness across varied garment styles.

Beyond Shop-to-Model scenarios, Tab. 3 systematically compares *cross-scenario* try-on performance. Missing entries (‘-’) denote scenario-specific limitations of certain methods¹. In Shop-to-Street tasks, D⁴-VTON and StreetTryOn, benefiting from warping priors, outperform prior-free methods, yet OmniVTON surpasses them in body reconstruction through Spectral Pose Injection (SPI). In Model-to-Model, Model-to-Street, and Street-to-Street settings, our training-free framework maintains a significant advantage, even outperforming StreetTryOn despite it being trained on in-domain data. Moreover, StreetTryOn struggles with lower-body garments and dresses in Shop-to-Street and Shop-to-Model tasks, as garment DensePose [10] fails to provide reliable predictions for these clothing categories. In contrast, SGM successfully generates pseudo-person images, ensuring comprehensive cross-scenario applicability.

¹We exclude GP-VTON due to unavailable parsing models.



Figure 4. Qualitative results across multiple datasets and clothing types. We provide upper garment try-on results on the VITON-HD dataset [8] (top) and lower garment/dresses visual comparisons on the DressCode dataset [29] (bottom).



Figure 5. Qualitative results on the StreetTryOn benchmark [11] under different scenarios (from top to bottom): Shop-to-Street, Model-to-Model, Model-to-Street, and Street-to-Street.

Qualitative Evaluation. We present qualitative results on the VITON-HD and DressCode datasets in Fig. 4. TIGIC

and Cross-Image fail to generate realistic human images due to their lack of task-specific designs. While inpainting models (PBE, AnyDoor) and traditional VTON methods improve human-body generation, they fail to transfer garment textures effectively and introduce noticeable artifacts, especially in cross-domain scenarios. As a GAN-based [17] method, GP-VTON exhibits poor human-body completion due to its limited generalization capability. In contrast, OmniVTON achieves high-fidelity try-on results while preserving garment textures with precision.

Since StreetTryOn has not released its code, we compare six alternative methods on this benchmark. Among them, CAT-DM and D⁴-VTON are restricted to in-shop garment inputs, while PWS and PastaGAN++ extract target garments from model images. As shown in Fig. 5, although generic inpainting models (PBE, AnyDoor) demonstrate adaptability to different scenarios, they fail to maintain pose and texture consistency between pre- and post-try-on images. Scenario-specific methods like PWS and PastaGAN++, trained on constrained backgrounds, struggle with real-world complexities. OmniVTON consistently achieves accurate garment texture transfer and pose alignment across diverse settings, demonstrating superior generalizability.

4.3. Ablation Analysis

We conduct an ablation analysis to assess the contributions of OmniVTON’s core modules through four model vari-

Method	SGM	CBS	SPI	FID _u ↓	FID _p ↓	SSIM _p ↑	LPIPS _p ↓
Base				18.445	16.878	0.773	0.222
(A)	✓			13.303	11.475	0.809	0.177
(B)	✓	✓		9.799	7.993	0.824	0.158
(C)			✓	13.148	10.767	0.813	0.180
OmniVTON	✓	✓	✓	9.621	7.758	0.832	0.145

Table 4. Ablation study of the Structured Garment Morphing (SGM), Continuous Boundary Stitching (CBS), and Spectral Pose Injection (SPI) on the VITON-HD dataset [8].



Figure 6. Qualitative ablation study on different variants.

ants. Starting from a baseline model (Base) using only text prompts, we incrementally integrate: (A) Structured Garment Morphing (SGM) for garment priors, (B) Continuous Boundary Stitching (CBS) for boundary refinement, and (C) Spectral Pose Injection (SPI) for pose-aware noise control. OmniVTON combines all components.

Effectiveness of SGM. As shown in Tab. 4, (A) significantly outperforms Base across all metrics, confirming that SGM effectively aligns garments without end-to-end training. Fig. 6 further illustrates its ability to preserve garment textures via fine-grained geometric cues.

Effectiveness of CBS. CBS eliminates boundary artifacts (red box, Fig. 6) and refines textures (blue box). Variant (B) improves LPIPS by 0.019 over (A), validating its role in enhancing perceptual quality. Since CBS primarily refines SGM-derived priors, we focus on their combined effect.

Effectiveness of SPI. As seen in Tab. 4, (C) and OmniVTON show notable SSIM and FID_u gains, confirming SPI’s ability to suppress noise contamination while preserving structural consistency. Fig. 6 (green circle) highlights improved body part alignment, reducing pose misalignment. Integrating all components, OmniVTON achieves state-of-the-art texture fidelity and pose consistency.

4.4. Multi-Human Virtual Try-On

Beyond single-human virtual try-on, our method extends to multi-human interactive group try-on (Fig. 7). This capability arises from the innovative design of SGM, which enables seamless adaptation for multiple users. Given multiple garments, we concatenate them along spatial dimensions to generate multiple pseudo-person images simulta-

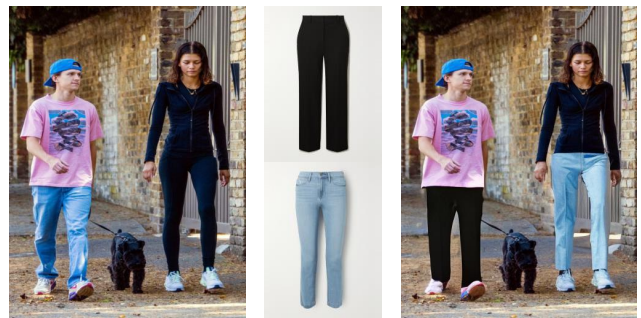


Figure 7. Multi-human virtual try-on. Top row shows Model-to-Model, while bottom row depicts Shop-to-Street.

neously. By leveraging positional and semantic cues from skeleton and parsing maps, our approach allows for the effortless application of identical or distinct garments to multiple humans. Multi-human try-on broadens the scope of virtual try-on tasks, further demonstrating the universality of our method. This extension opens new directions for group-centric fashion experiences, such as coordinated family outfits and uniform design.

5. Conclusions, Limitations, and Future Work

In this paper, we present OmniVTON, a training-free universal framework that ensures both texture fidelity and pose consistency across diverse settings. Structured Garment Morphing enables anatomical garment-body alignment, while Continuous Boundary Stitching ensures seamless texture transitions, achieving fine-grained texture consistency without domain-specific training. Spectral Pose Injection further enhances pose alignment through frequency-aware modulation of inversion noise, preserving structural integrity while eliminating texture contamination. Extensive experiments demonstrate OmniVTON’s superiority in flexibility and generalization, particularly in its pioneering capability for multi-human VTON.

Despite its effectiveness, OmniVTON encounters challenges in extreme cases, such as high-density crowds or minimal target body regions, leading to garment misalignment. Visual results illustrating these limitations are provided in the supplementary materials. Future work will focus on developing more robust multi-human try-on frameworks to address these edge cases.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 62102381, 41927805); Shandong Natural Science Foundation (No. ZR2021QF035); the National Key R&D Program of China (No. 2022ZD0117201); the Guangdong Natural Science Funds for Distinguished Young Scholar (No.2023B1515020097); the National Research Foundation, Singapore under its AI Singapore Programme (No.AISG3-GV-2023-011); the Singapore Ministry of Education AcRF Tier 1 Grant (No. MSS25C004); and the Lee Kong Chian Fellowships.

References

- [1] Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. Cross-image attention for zero-shot appearance transfer. In *ACM SIGGRAPH*, pages 1–12, 2024. 5, 6
- [2] Badour AlBahar, Jingwan Lu, Jimei Yang, Zhixin Shu, Eli Shechtman, and Jia-Bin Huang. Pose with style: Detail-preserving pose-guided image synthesis with conditional stylegan. *ACM TOG*, 40(6):1–11, 2021. 5, 6
- [3] Moab Arar, Rinon Gal, Yuval Atzmon, Gal Chechik, Daniel Cohen-Or, Ariel Shamir, and Amit H. Bermano. Domain-agnostic tuning-encoder for fast personalization of text-to-image models. In *SIGGRAPH Asia*, pages 1–10, 2023. 3
- [4] Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE TPAMI*, 11(6):567–585, 1989. 2
- [5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, pages 7291–7299, 2017. 2, 4
- [6] 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. In *ECCV*, pages 124–142. Springer, 2024. 3
- [7] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *CVPR*, pages 6593–6602, 2024. 3, 5, 6
- [8] 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. 2, 4, 5, 6, 7, 8
- [9] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for authentic virtual try-on in the wild. In *ECCV*, pages 206–235. Springer, 2024. 3, 5, 6
- [10] Aiyu Cui, Sen He, Tao Xiang, and Antoine Toisoul. Learning garment densepose for robust warping in virtual try-on. *arXiv preprint arXiv:2303.17688*, 2023. 6
- [11] Aiyu Cui, Jay Mahajan, Viraj Shah, Preeti Gomathinayagam, Chang Liu, and Svetlana Lazebnik. Street tryon: Learning in-the-wild virtual try-on from unpaired person images. In *WACV*, pages 8235–8239, 2025. 2, 3, 5, 6, 7
- [12] Daan De Geus and Gijs Dubbelman. Task-aligned part-aware panoptic segmentation through joint object-part representations. In *CVPR*, pages 3174–3183, 2024. 4
- [13] Yong Du, Jiahui Zhan, Xinzhe Li, Junyu Dong, Sheng Chen, Ming-Hsuan Yang, and Shengfeng He. One-for-all: towards universal domain translation with a single stylegan. *IEEE TPAMI*, 47(4):2865–2881, 2025. 2
- [14] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 3
- [15] Henri P Gavin. The levenberg-marquardt algorithm for non-linear least squares curve-fitting problems. *Department of Civil and Environmental Engineering Duke University August*, 3:1–23, 2019. 4
- [16] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *CVPR*, pages 5337–5345, 2019. 5
- [17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 7
- [18] 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. In *ACM MM*, pages 7599–7607, 2023. 2
- [19] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, pages 7297–7306, 2018. 2
- [20] 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
- [21] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *ICCV*, pages 10471–10480, 2019. 2
- [22] Yucheng Han, Rui Wang, Chi Zhang, Juntao Hu, Pei Cheng, Bin Fu, and Hanwang Zhang. Emma: Your text-to-image diffusion model can secretly accept multi-modal prompts. *arXiv preprint arXiv:2406.09162*, 2024. 4
- [23] Sen He, Yi-Zhe Song, and Tao Xiang. Style-based global appearance flow for virtual try-on. In *CVPR*, pages 3470–3479, 2022. 2
- [24] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *ECCV*, pages 150–168. Springer, 2024. 3
- [25] Jeongho Kim, Guojung Gu, Minhoo Park, Sunghyun Park, and Jaegul Choo. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *CVPR*, pages 8176–8185, 2024. 3
- [26] 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

- [27] Pengzhi Li, Qiang Nie, Ying Chen, Xi Jiang, Kai Wu, Yuhuan Lin, Yong Liu, Jinlong Peng, Chengjie Wang, and Feng Zheng. Tuning-free image customization with image and text guidance. In *ECCV*, pages 233–250. Springer, 2024. 3, 5, 6
- [28] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, pages 1096–1104, 2016. 6
- [29] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: High-resolution multi-category virtual try-on. In *CVPR*, pages 2231–2235, 2022. 5, 6, 7
- [30] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. In *ACM MM*, pages 8580–8589, 2023. 3
- [31] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, pages 11410–11420, 2022. 6
- [32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 3
- [33] Haorui Song, Yong Du, Tianyi Xiang, Junyu Dong, Jing Qin, and Shengfeng He. Editing out-of-domain gan inversion via differential activations. In *ECCV*, pages 1–17. Springer, 2022. 2
- [34] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ICLR*, 2020. 2, 4
- [35] Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH*, pages 1–11, 2023. 4
- [36] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 6
- [37] Zhenyu Xie, Zaiyu Huang, Fuwei Zhao, Haoye Dong, Michael Kampffmeyer, and Xiaodan Liang. Towards scalable unpaired virtual try-on via patch-routed spatially-adaptive gan. *NeurIPS*, 34:2598–2610, 2021. 6
- [38] Zhenyu Xie, Zaiyu Huang, Fuwei Zhao, Haoye Dong, Michael Kampffmeyer, Xin Dong, Feida Zhu, and Xiaodan Liang. Pasta-gan++: A versatile framework for high-resolution unpaired virtual try-on. *arXiv preprint arXiv:2207.13475*, 2022. 2, 5, 6
- [39] 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, 4, 5, 6
- [40] Yuhao Xu, Tao Gu, Weifeng Chen, and Chengcai Chen. Oot-diffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. *arXiv preprint arXiv:2403.01779*, 2024. 2, 3
- [41] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *CVPR*, pages 18381–18391, 2023. 3, 5, 6
- [42] Zhaotong Yang, Zicheng Jiang, Xinzhe Li, Huiyu Zhou, Junyu Dong, Huaidong Zhang, and Yong Du. D⁴-vton: Dynamic semantics disentangling for differential diffusion based virtual try-on. In *ECCV*, pages 36–52. Springer, 2024. 2, 5, 6
- [43] Jianhao Zeng, Dan Song, Weizhi Nie, Hongshuo Tian, Tongtong Wang, and An-An Liu. Cat-dm: Controllable accelerated virtual try-on with diffusion model. In *CVPR*, pages 8372–8382, 2024. 3, 5, 6
- [44] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 6
- [45] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *ECCV*, pages 286–301. Springer, 2016. 2
- [46] Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. In *ECCV*, pages 195–211. Springer, 2024. 3