FitControler: Toward Fit-Aware Virtual Try-On

Lu Yang[1], Yicheng Liu[1], Letian Zhou[1], Yanan Li[2], Xiang Bai[1], and Hao Lu[1,⋆]

[1] Huazhong University of Science and Technology, China
[2] Wuhan Institute of Technology, China
{lu_yang1, hlu}@hust.edu.cn

Figure 1: Illustrations of FitControler generations. Zoom in and compare contour differences of the garment fit.

Abstract Realistic virtual try-on (VTON) concerns not only faithful rendering of garment details but also coordination of the style. Prior art typically pursues the former, but neglects a key factor that shapes the holistic style—garment fit. Garment fit delineates how a garment aligns with the body of a wearer and is a fundamental element in fashion design. In this work, we introduce fit-aware VTON and present FitControler, a learnable plug-in that can seamlessly integrate into modern VTON models to enable customized fit control. To achieve this, we highlight two challenges: i) how to delineate layouts of different fits and ii) how to render the garment that matches the layout. FitControler first features a fit-aware layout generator to redraw the body-garment layout conditioned on a set of delicately processed garment-agnostic representations, and a multi-scale fit injector is then used to deliver layout cues to enable layout-driven VTON. In particular, we build a fit-aware VTON dataset termed Fit4Men, including 13,000 body-garment pairs of different fits, covering both tops and bottoms, and featuring varying camera distances and body poses. Two fit consistency metrics are also introduced to assess the fitness of generations. Extensive experiments show that FitControler can work with various VTON models and achieve accurate fit control. Code and data are released at `https://github.com/tiny-smart/FitControler`.

Keywords: Fit Control · Virtual Try-On · Diffusion Models

⋆ Corresponding author.

1 Introduction

"Does this outfit suit me?"—this is a common question for online shoppers, yet difficult to answer through imagination alone. VTON seeks to bridge this gap by generating realistic try-on images of a wearer with desired garments. Realistic VTON generation, however, is difficult. It hinges on not only faithful preservation of garment appearance and body pose but also coordination of the holistic style. Existing work [4–6, 17, 36, 50] mainly focuses on recovering garment details while overlooking the style consistency. Per Fig. 2(a), this negligence can result in disharmony of the feeling and may mismatch user preferences.

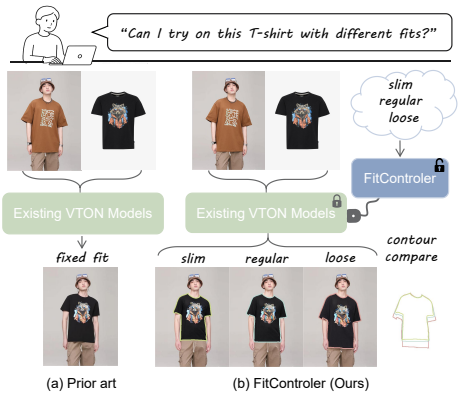

Figure 2: Comparison between existing VTON models and our proposed FitControler. (a) Existing models produce only a fixed fit, often leading to unnatural results due to mismatched fit. (b) With FitControler, the same inputs can produce try-on results with customized fits such as *slim*, *regular*, and *loose* T-shirts, which much better matches the overall style and user preference.

In the fashion domain, style is primarily influenced by three factors: color coordination, ways of wearing, and fit [20, 45]. Since the garment is preset, the latter two matter in VTON. Ways of wearing—such as tucking in the hem or rolling up sleeves—affect local styles, while *fit shapes the holistic style*, delineates how a garment aligns with the body, and is governed by physical factors such as garment proportions, measurements, and ease allowance. These physical properties manifest visually via variations in tightness (*e.g.*, slim vs. loose T-shirts) and shape (*e.g.*, tapered vs. straight trousers) [45], making it possible to reproduce target fits visually without the need for precise physical modeling.

In this work, we introduce fit-aware VTON and present FitControler, a novel plug-in engineered for seamless integration with modern diffusion-based VTON models, empowering users with customized fit control (Fig. 2(b)). We remark that a key challenge for fit-aware VTON lies in *how to ground the abstract notion of fit into tangible visual representations*. To this end, FitControler follows a two-stage design: i) delineating the spatial layout between the garment and body given a certain fit, and ii) rendering the garment onto the body conditioned on the layout.

The first stage deals with spatial modeling between the body and garment. Following [4, 44], we use the segmentation map to represent the body-garment layout and introduce a fit-aware layout generator to redraw the target segmentation map conditioned on the fit label. During layout generation, a phenomenon called *garment shape leakage* arises. This leads to a problem where the generator is biased toward replicating the original garment shape, instead of generating a

new one. To address this, we redesign the garment-agnostic representation preprocessor to produce a rectangular mask and a dense pose conditioned on body proportions. The second stage follows a conditional image generation pipeline. Unlike previous approaches [27,46] that often rely on an extra encoder to extract conditional image features, we repurpose the features from the layout generator, and a lightweight multi-scale fit injector is used to deliver the layout features to off-the-shelf VTON models via the ControlNet [46] interface. We find that such a simplification not only reduces the computational overhead but also accelerates training convergence.

In particular, we harvest a fit-orientated VTON dataset, termed Fit4Men. We focus on men garments because they feature limited but well-defined garment styles compared with more varied women garments, making them suitable for an initial exploration on this topic. Fit4Men features $5,000$ pairs of men T-shirts with `slim`, `regular`, and `loose` fits and $8,000$ pairs of men trousers with `tapered` and `straight` fits. The dataset covers diverse body poses and varying camera distances. To objectively assess the fitness of try-on generations, two fit consistency metrics measuring contour differences are also introduced.

Experiments demonstrate that FitControler can be incorporated into various VTON architectures, providing precise control over garment fit while enhancing the perceptual quality of try-on generations. Ablation studies further show that effective fit control can be achieved with as few as 1,000 training steps or with only 1,000 training samples, highlighting good training and sample efficiency.

2 Related Work

Our work is related to image-based VTON, customized VTON, and controllable image generation.

Image-based Virtual Try-On. Image-based VTON aims to transfer a target garment onto a person and synthesize photorealistic try-on images. Conventional methods [4, 11, 22, 37, 44] typically follow a two-stage pipeline by i) warping the garment to align with the target human pose and ii) compositing try-on images using a generative model [8]. These approaches, however, can suffer from poor geometric warping under large deformations and unrealistic garment rendering due to the limited generation quality.

Recent diffusion-based methods commonly leverage attention mechanisms to model garment-person interactions, reducing reliance on explicit warping. They differ mainly in how garment details are captured. For instance, LaDI-VTON [26] and DCI-VTON [9] encode garments using CLIP [34], but CLIP embeddings fail to preserve fine-grained textures due to their focus on high-level semantics. To address this, StableVITON [17] employs a ControlNet-like [46] architecture to capture garment features, enhancing detail preservation. Following TryOnDiffusion [52], mainstream approaches [5, 15, 36, 40, 50] repurpose the denoising backbone to extract garment features, further improving texture fidelity. Although effective, this dual-branch design significantly increases training and inference

costs. TPD [42] and CatVTON [6] streamline the pipeline by jointly encoding garment and person features within a single U-Net, boosting both efficiency and fidelity. Despite these advances, they overlook the wearing style, often leading to unnatural results. In particular, the inconsistency of garment fit—such as mismatched tightness between tops and bottoms—can significantly degrade realism. To address this, we explicitly model the garment fit to enable fit-aware VTON.

Customized Virtual Try-On. Due to personalized demands, some work has also explored customized VTON. Landmark-based methods [2, 3, 23, 41] enable control over local wearing styles (e.g., sleeve rolling or hem tucking), but typically require manual landmark adjustment. COTTON [2] removes manual intervention and supports garment length editing, yet overlooks ease allowance, which is critical for proper fit. Text-based approaches [18, 24, 51] synthesize try-on images from textual descriptions, offering potential attribute control. PromptDresser [18] demonstrates impressive text-driven editing capabilities; however, its control over garment fit remains implicitly entangled with language semantics and pixel-level inpainting, lacking explicit modeling of body–garment spatial alignment. Furthermore, the controllability of these approaches is mainly applicable to specific model architectures—even text-based control cannot be applied to all diffusion-based models (*e.g.*, CatVTON [6] drops text input). In this work, we consider garment fit as a holistic, spatially grounded attribute and design a dedicated plug-in for modern diffusion-based VTON models.

Controllable Image Generation. Recently, notable progress in controllable image generation [33, 39, 49] has enabled conditional guidance such as segmentation maps, canny edges, or depth maps, without retraining text-to-image diffusion models [28, 29, 35]. For example, ControlNet [46] only trains an auxiliary encoder mirroring the structure of the denoising model, and T2I-Adapter [27] introduces a lightweight network for the same purpose. These methods, however, focus solely on guiding generation from a conditional image, without considering how the condition is acquired. As a result, they typically feature an additional encoder to process the conditional input. In contrast, our approach generates the conditional image itself, allowing us to repurpose features from the generator, which simplifies model design and improves efficiency.

3 FitControler

We begin with the problem setup and then present the key technical details of FitControler.

3.1 Problem Setup

Given a person image $\boldsymbol{x}_p$ and a garment image $\boldsymbol{x}_g$, fit-aware VTON aims to synthesize a try-on image $\boldsymbol{x}_{tr}$ according to a target fit prompt $\boldsymbol{l}$, where $\boldsymbol{x}_p, \boldsymbol{x}_g, \boldsymbol{x}_{tr} \in \mathbb{R}^{H\times W\times 3}$, with H being the image height and W the image width, and $\boldsymbol{l}$ can be of various forms, such as text or categorical labels.

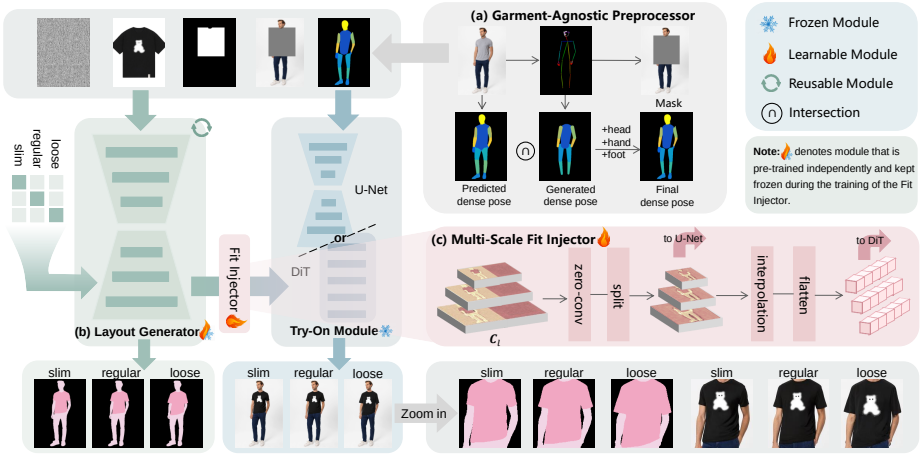

Figure 3: Overview of FitControler. The person image is first processed by (a) the *garment-agnostic preprocessor* to extract the mask and dense pose. These are concatenated with the noise map and garment image—along both channel and spatial dimensions as is done in CatVTON [6]—before being fed into (b) the *fit-aware layout generator* to produce a fit-sensitive segmentation map. The layout features are then delivered by (c) the *multi-scale fit injector* to the base VTON models via the ControlNet [46] interface. The layout generator is pretrained and remains frozen when integrating FitControler into different VTON models, where only the fit injector requires model-specific training.

A common paradigm [3, 18] is to design a dedicated try-on model $\mathcal{T}$ that accepts $\boldsymbol{l}$ as an additional input such that

$$\boldsymbol{x}_{tr} = \mathcal{T}(\mathcal{P}(\boldsymbol{x}_p), \boldsymbol{x}_g, \boldsymbol{l}), \tag{1}$$

where $\mathcal{P}(\boldsymbol{x}_p)$ is a preprocessor that generates garment-agnostic representations (*e.g.*, masked images) [4] from $\boldsymbol{x}_p$. This paradigm, however, tightly couples the fit control mechanism with a certain model and cannot easily integrate into other VTON models.

Unlike the common paradigm, our idea is to design a plug-in $\mathcal{F}$ that is compatible with existing VTON models. This requires addressing two key problems: i) how to encode the fit prompt $\boldsymbol{l}$ into the fit feature $\boldsymbol{C}_l$ with clear fit awareness and ii) how to enable try-on models to effectively decode $\boldsymbol{C}_l$. We formulate the two problems as

$$\boldsymbol{C}_l = \mathcal{F}(\mathcal{P}(\boldsymbol{x}_p), \boldsymbol{x}_g, \boldsymbol{l}), \tag{2}$$

$$\boldsymbol{x}_{tr} = \mathcal{T}(\mathcal{P}(\boldsymbol{x}_p), \boldsymbol{x}_g, \boldsymbol{C}_l). \tag{3}$$

3.2 FitControler Overview

Fig. 3 shows the technical pipeline of FitControler. FitControler instructs the fit with categorical labels and features i) a *garment-agnostic preprocessor* utilizing

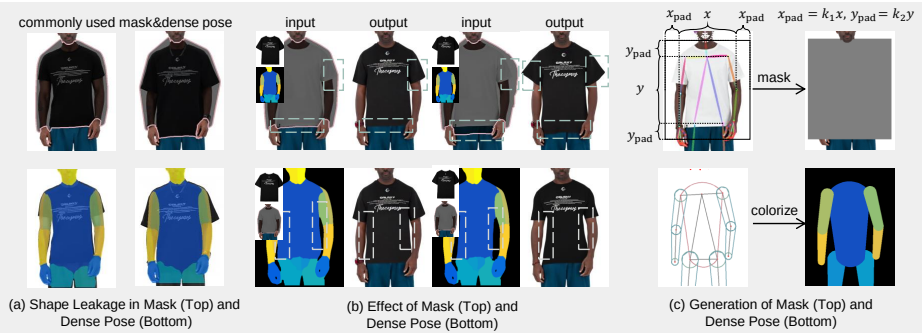

Figure 4: Impact of shape leakage in mask and dense pose. (a) Commonly used masks and dense poses implicitly encode the original garment boundaries, which (b) biases the model to follow these cues when rendering garment. To address this, (c) our preprocessor reconstructs them from human keypoints to form standardized representations. See supplementary material for the ablation on this preprocessor.

human keypoints and body proportions to generate a square-shaped mask and a simulated dense pose; ii) a *fit-aware layout generator* producing the fit feature that encodes the body-garment spatial relation, conditioned on the fit label; and iii) a *multi-scale fit injector* mapping the fit feature compatible with try-on models and injects it via the ControlNet [46] interface.

3.3 Garment-Agnostic Preprocessor

Garment-agnostic representations are standard inputs to VTON models that aim to reduce the cost of data collection. They enable training with (`garment`, `try-on image`) pairs under reconstruction-based supervision. However, commonly used representations such as the mask and the dense pose (Fig. 4(a)) often preserve the original garment contour, because the mask is typically generated along the garment boundary, and DensePose [10] tends to misinterpret garment edges as body boundaries. Consequently, the model is biased to render garments following these leaked cues (Fig. 4(b)), indicating that the standard representations do not meet the need of fit-aware synthesis. To address this, we redesign the garment-agnostic preprocessor $\mathcal{P}(\boldsymbol{x}_p)$ to eliminate garment shape leakage.

Mask. Unlike prior work [4] that estimates masks from detected garment regions, we derive masks from human keypoints to avoid garment contour leakage. Specifically, for the upper body, we define x as the horizontal span between the minimum and maximum coordinates of the shoulder, elbow, and wrist joints, and y as the vertical span between the shoulder and hip keypoints. The mask is generated as in Fig. 4(c), where two empirical padding ratios $k_1 = 0.6$ and $k_2 = 0.25$ to fully cover the clothing region while minimizing unnecessary background. For the lower body, we construct the mask based on the hip, knee, and ankle keypoints, with $k_1 = 0.5$ and $k_2 = 0.2$.

DensePose. To mitigate inaccuracies in the predicted dense pose, we synthesize a standard-stature dense pose based on human keypoints and canonical body proportions. According to Fig. 4(c), the torso is approximated by a quadrilateral defined by shoulder and hip joints, with convex arcs at the top and bottom to yield natural contours. Limbs are constructed by connecting circles placed at major joints (shoulder–elbow–wrist and hip–knee–ankle), whose diameters are proportional to body height (*e.g.*, 0.06, 0.048, 0.033 for the arm; 0.09, 0.055, 0.03 for the leg). Body height is estimated to be of 3.2× torso height or 2.3× leg length. Finally, the synthesized map is intersected with the predicted dense pose to constrain the spatial layout.

Notably, what matters here is the general design principle of constructing garment-agnostic masks and dense poses to eliminate shape leakage; the specific empirical parameters are not critical to the overall effectiveness.

3.4 Fit-Aware Layout Generator

The fit-aware layout generator produces a body-garment layout conditioned on the fit label $\boldsymbol{l}$. We represent the layout as a segmentation map and repurpose the U-Net from Stable Diffusion [35] to achieve this task. The rich semantic priors from the pretrained U-Net can benefit layout prediction.

Formally, given a person image $\boldsymbol{x}_p$ and a garment image $\boldsymbol{x}_g$, we first derive the mask $\boldsymbol{m}$, the masked person image $\boldsymbol{x}_m$, and the dense pose $\boldsymbol{x}_d$ via $\mathcal{P}(\boldsymbol{x}_p)$. Since the U-Net operates in the latent space of a VAE [19], these inputs are first encoded by the VAE encoder $\mathcal{E}(\cdot)$, with the mask $\boldsymbol{m}$ downsampled to the corresponding resolution. A Gaussian noise map $\boldsymbol{z}_T$ is further sampled as the stochastic input. The input to the generator is a 13-channel tensor

$$\mathcal{X} = \boldsymbol{z}_T \,||\, [\boldsymbol{m} \oplus \mathbf{0}] \,||\, [\mathcal{E}(\boldsymbol{x}_m) \oplus \mathcal{E}(\boldsymbol{x}_g)] \,||\, [\mathcal{E}(\boldsymbol{x}_d) \oplus \mathbf{0})]\,, \tag{4}$$

along with the one-hot encoded fit label $\boldsymbol{l}$, where $\mathbf{0}$ is a zero tensor used for spatial alignment, $\oplus$ denotes spatial concatenation, and $||$ channel-wise concatenation, following CatVTON [6]. Then the layout generator $\mathcal{G}_{\mathcal{S}}$ produces

$$\boldsymbol{S}_l, \boldsymbol{C}_l = \mathcal{G}_{\mathcal{S}}(\mathcal{X}, \boldsymbol{l})\,, \tag{5}$$

where $\boldsymbol{S}_l$ is a three-class segmentation map (`background`, `body`, and `garment`), and $\boldsymbol{C}_l$ is the fit-aware features aggregated from multi-scale intermediate decoder features before ResNet blocks. To condition $\boldsymbol{l}$, we follow the prior controlled image generation pipeline [29,32] and inject it through feature-level modulation using FiLM [30] in each ResNet block of the decoder, which amounts to

$$\boldsymbol{h}_s \leftarrow \gamma(\boldsymbol{l}) \cdot \frac{\boldsymbol{h}_s - \mu}{\sigma} + \beta(\boldsymbol{l})\,, \tag{6}$$

where $\boldsymbol{h}_s$ denotes intermediate features, μ and σ are feature-wise statistics, and $\gamma(\cdot)$ and $\beta(\cdot)$ are learnable projections. Notably, we freeze the U-Net encoder to preserve its pretrained semantic knowledge. Additional implementation details are provided in the supplementary material.

3.5 Multi-Scale Fit Injector

The fit injector transforms the multi-scale layout features $\boldsymbol{C}_l$ into a compatible representation that is acceptable by VTON models.

At each scale, a zero-initialized convolution layer is first applied to align the layout features with try-on features. To ensure spatial consistency, we apply operations such as splitting and interpolation to match the resolution of the try-on model. For U-Net based models (*e.g.*, Leffa [50]), resolution alignment can be achieved by simple splitting operations. In contrast, DiT-based models (*e.g.*, FitDiT [15]) require interpolated feature maps of a uniform resolution, followed by flattening.

During the training of the injector, both the layout generation module and the try-on module remain frozen, and only the injector parameters are optimized using the loss

$$\mathcal{L} = \mathbb{E}_{\zeta,t,\epsilon,C_l}\left[\|\epsilon - \epsilon_\theta(\boldsymbol{\zeta}, t, \tau_\phi(\boldsymbol{C}_l))\|_2^2\right], \tag{7}$$

where τ_ϕ is the injector, t is the diffusion timestep, ϵ is the noise, and $\boldsymbol{\zeta}$ denotes the try-on input. For example, in Leffa, $\boldsymbol{\zeta} = [\mathcal{E}(\boldsymbol{x}_g),\ \boldsymbol{z}_t \,||\, \mathcal{E}(\boldsymbol{x}_m) \,||\, \mathcal{E}(\boldsymbol{x}_d) \,||\, \boldsymbol{m}]$.

4 Fit Consistency Metric

Evaluating how well a try-on image adheres to a specific fit is non-trivial, because fit differences manifest in not only realism but also subtle garment contours. This requires explicit contour-aware metrics to compare fit consistency. Concretely, for each source image, we generate a try-on image conditioned on the same fit label, extract the garment contours from both, and measure their similarity. We adopt two complementary shape metrics: Hu moments (Hu) [13] and Hausdorff distance (Hd) [14].

Hu Moments. Hu delineates the shape globally. Given a binary contour mask $I(x, y)$, the central moment of order (p, q) is first computed by

$$\mu_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q I(x, y)\,, \tag{8}$$

where $(\bar{x}, \bar{y})$ is the centroid. It is then normalized to

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^r}, \quad r = \frac{p+q}{2} + 1\,. \tag{9}$$

Hu takes the form $\Phi(I) = [\phi_1, \ldots, \phi_7]$, where each ϕ_i is a specific combination of η_{pq}'s, *e.g.*, $\phi_1 = \eta_{20} + \eta_{02}$. Hu captures the spatial distribution of contour pixels via moment statistics, providing a compact yet informative representation of global shape. In practice, we compute the Hu distance between two images as a single scalar. The closer the distance between $\Phi(I_{\text{gen}})$ and $\Phi(I_{\text{src}})$ is, the better the garment contour matches, where I_{gen} and I_{src} refer to the generated and source contours, respectively.

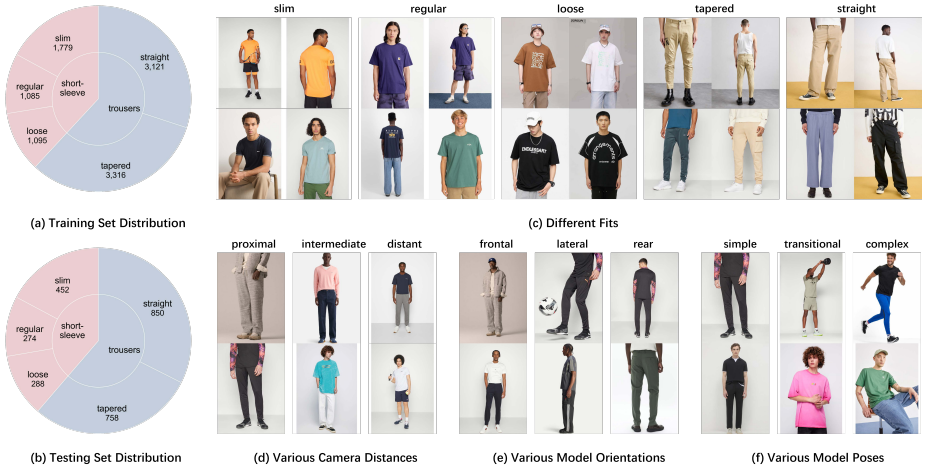

Figure 5: Overview of Fit4Men. (a) and (b) present the statistical distributions of the training and testing sets. (c) Examples of five garment fit types supporting fit-aware learning. (d) shows samples captured at different camera distances, while (e) and (f) depict variations in model orientations and poses, respectively, ensuring the diversity of Fit4Men.

Hausdorff Distance. Hd captures local geometric deviations. Given two contour point sets A and B, the Hausdorff distance $d_H(A, B)$ between A and B takes the form

$$d_H(A, B) = \max\left\{\sup_{a\in A}\inf_{b\in B} d(a, b),\ \sup_{b\in B}\inf_{a\in A} d(b, a)\right\}, \tag{10}$$

where $d(a, b)$ is the Euclidean distance between points a and b in our case. Hd measures the maximum nearest-neighbor distance between contour points, highlighting mismatches such as the misalignment of sleeve or hem.

With Hu and Hd, our fit consistency metrics take both global shape coherence and local fidelity into account. Lower values of both metrics indicate better performances.

5 Fit4Men Dataset

We construct **Fit4Men**, a medium-resolution (768×1024) dataset specifically designed for fit-aware VTON. As illustrated in Fig. 5, it comprises 5,000 pairs of men's short-sleeve shirts categorized into three fit types (`slim`, `regular`, and `loose`) along with 8,000 pairs of men's trousers featuring two fit types (`tapered` and `straight`) sourced from e-commerce platforms.[3] The dataset is split into training and testing sets at a ratio of 4:1. Regarding annotations, human keypoints are extracted using DWPose [43], while DensePose [10] and Sapiens [16]

[3] Zalando, Taobao, and Musinsa.

are employed to generate dense poses and ground-truth segmentation maps, respectively. To replicate realistic try-on scenarios, Fit4Men incorporates variations in camera distance, model orientation, and pose complexity.

Camera Distance. Fit4Men encompasses a broad range of camera distances frequently encountered in fashion photography. These distances are classified into three levels consisting of proximal, intermediate, and distant views. Proximal views capture specific body regions (*e.g.*, upper or lower body), whereas intermediate views depict the full body within the frame, and distant views portray the subject with significant spatial separation from the camera. This variation enhances robustness against framing discrepancies.

Model Orientation. The dataset features subjects captured from diverse orientations, specifically frontal, lateral, and rear views. Such viewpoint diversity mirrors real-world catalog photography while enhancing generalization to arbitrary viewing angles during the virtual try-on process.

Pose Complexity. Fit4Men addresses three levels of pose complexity including simple, moderate, and complex categories. Simple poses represent upright standing positions characterized by an absence of limb occlusion. Moderate poses involve partial self-occlusion (*e.g.*, crossed arms or bent legs). Complex poses encompass dynamic or non-standing states including sitting or running. This design facilitates the evaluation of fit consistency under diverse body configurations.

6 Results and Discussion

Here we present our results, and discussion.

6.1 Experimental Setup

Implementation Details. We first train the layout generator using the cross-entropy loss. Subsequently, for each try-on model, the fit injector is trained for 7,500 steps following the configuration described in Sec. 3.5. All models are optimized using AdamW [25] with a batch size of 64 and a learning rate of 1×10^{-5} on Fit4Men. Training and evaluation are performed at 384×512 resolution with FP16 precision on four NVIDIA A6000 GPUs. Additional training details are provided in the supplementary material.

Experimental Protocol. To demonstrate generality, we integrate FitControler into five state-of-the-art VTON models: StableVITON [17], IDM-VTON [5], CatVTON (and its FLUX variant) [6], Leffa [50], and FitDiT [15], covering diffusion backbones from SD1.5 [35] and SDXL [31] to SD3 [7] and FLUX.1 [21]. Each model is evaluated on Fit4Men with and without FitControler, under both paired and unpaired settings. The baseline VTON models use their developed preprocessing pipelines for generating garment-agnostic representations. Following prior work [6,48], we also report FID [12], KID [1], SSIM [38], and LPIPS [47] to assess image quality in the paired setting, and FID/KID in the unpaired setting. Both settings also report Hu and Hd to evaluate fit consistency.

Table 1: Analysis of fit consistency metrics. **pd** is the generated image, and **gt** the source image. Best performance is in **boldface**.

(a) Upper-body garments.

pd \ gt	slim		regular		loose	
	Hu	Hd	Hu	Hd	Hu	Hd
slim	**0.32**	**6.46**	0.43	7.60	0.67	14.34
regular	0.41	7.65	**0.39**	**6.35**	0.54	11.16
loose	0.58	11.61	0.47	9.24	**0.43**	**7.34**

(b) Lower-body garments.

pd \ gt	tapered		straight	
	Hu	Hd	Hu	Hd
tapered	**0.62**	**6.43**	1.74	13.14
straight	1.47	12.83	**0.91**	**9.47**

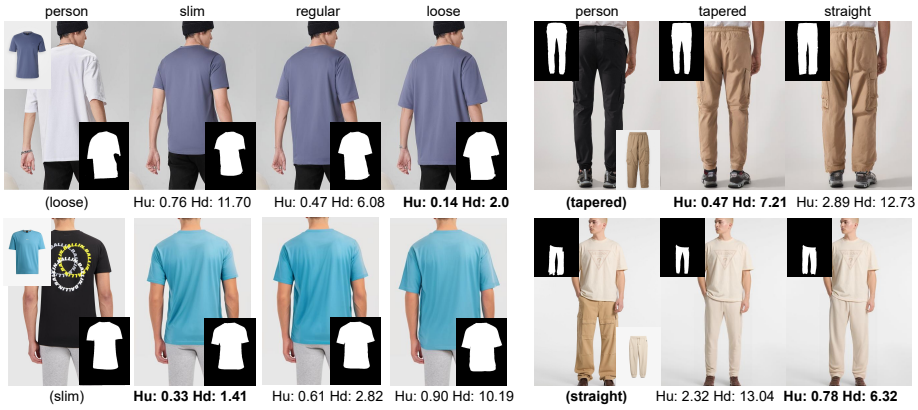

Figure 6: Correlation analysis of fit consistency metrics. For each target fit, three types of fits (slim, regular, and loose) are generated and compared.

6.2 Sanity Check on Fit Consistency Metrics

Before reporting Hu and Hd, we first perform a sanity check to confirm their soundness in assessing fit consistency. We analyze both qualitative and quantitative results on Fit4Men. For the qualitative analysis, we generate three different fits from a person-garment pair and compute the Hu and Hd values relative to the original image for short-sleeve shirts. For the quantitative analysis, we partition the the short-sleeve test set into three subsets according to their fit labels and generate VTON results w.r.t. the three different fits for each subset. An analogous protocol is applied to trousers. According to Table 1 and Fig. 6, both metrics achieve the best values when the generated fit matches the original one, and increase progressively with larger fit differences (*e.g.*, loose vs. slim). This confirms Hu and Hd are suitable for evaluating fit consistency.

6.3 Main Results

Here we compare the performance of state-of-the-art VTON models with and without FitControler on the Fit4Men dataset. Since StableVITON [17] provides pretrained weights only for upper-body data, it is only evaluated on the short-sleeve subset. Quantitative results are shown in Table 2. One can observe that FitControler consistently improves the fit consistency metrics in both Hu and

Table 2: Performance of VTON models with/without FitControler on Fit4Men. Relative improvements are colored in red ↑ and green ↓.

Method	Paired						Unpaired			
	FID↓	KID↓	SSIM↑	LPIPS↓	Hu↓	Hd↓	FID↓	KID↓	Hu↓	Hd↓
Short-sleeve										
StableVITON [17]	16.93	4.34	0.853	0.0653	0.57	9.90	21.97	5.75	0.66	10.35
+FitControler	**14.39** (15.0%)	**2.53** (41.7%)	**0.874** (2.5%)	**0.0634** (2.9%)	**0.42** (26.3%)	**7.63** (22.9%)	**19.96** (9.1%)	**4.81** (16.3%)	**0.50** (24.2%)	**8.54** (17.5%)
IDMVTON [5]	15.99	2.90	0.852	0.0647	0.55	9.03	20.83	3.37	0.67	10.39
+FitControler	**14.69** (8.1%)	**1.92** (33.8%)	**0.866** (1.6%)	**0.0592** (8.5%)	**0.45** (18.2%)	**7.54** (16.5%)	**19.11** (8.3%)	**2.19** (35.0%)	**0.51** (23.9%)	**8.81** (15.2%)
CatVTON [6]	14.28	2.02	0.864	0.0574	0.57	9.17	19.09	3.35	0.64	10.58
+FitControler	**13.01** (8.9%)	**1.06** (47.5%)	**0.873** (1.0%)	**0.0552** (3.8%)	**0.44** (22.8%)	**7.13** (22.2%)	**18.29** (4.2%)	**2.07** (38.2%)	**0.49** (23.4%)	**8.64** (18.3%)
CatVTON-FLUX [6]	15.76	2.54	0.855	0.0764	0.56	9.41	20.40	3.04	0.65	10.56
+FitControler	**14.44** (8.4%)	**1.39** (45.3%)	**0.860** (0.6%)	**0.0652** (14.7%)	**0.44** (21.4%)	**7.92** (15.8%)	**18.40** (9.8%)	**1.93** (36.5%)	**0.47** (27.7%)	**8.47** (19.8%)
FitDiT [15]	17.40	4.17	0.864	0.0723	0.65	12.36	23.94	6.82	0.80	16.89
+FitControler	**12.51** (28.1%)	**0.90** (78.4%)	**0.875** (1.3%)	**0.0510** (29.5%)	**0.38** (41.5%)	**6.77** (45.2%)	**18.48** (22.8%)	**1.86** (72.7%)	**0.46** (42.5%)	**7.94** (53.0%)
Leffa [50]	14.59	1.03	0.861	0.0668	0.56	9.18	18.77	2.59	0.65	10.47
+FitControler	**12.45** (14.7%)	**0.62** (39.8%)	**0.869** (0.9%)	**0.0530** (20.7%)	**0.40** (28.6%)	**7.05** (23.2%)	**17.73** (5.5%)	**1.54** (40.5%)	**0.45** (30.8%)	**7.65** (26.9%)
Trousers										
IDMVTON [5]	15.11	6.67	0.854	0.0677	1.03	10.53	16.12	5.59	1.58	12.58
+FitControler	**12.14** (19.7%)	**3.68** (44.8%)	**0.862** (0.9%)	**0.0592** (12.6%)	**0.81** (21.4%)	**9.32** (11.5%)	**14.43** (10.5%)	**3.95** (29.3%)	**1.13** (28.5%)	**10.29** (18.2%)
CatVTON [6]	11.46	3.43	0.852	0.0637	0.92	9.50	14.19	3.94	1.48	11.92
+FitControler	**10.02** (12.6%)	**1.82** (46.9%)	**0.867** (1.8%)	**0.0582** (8.6%)	**0.80** (13.0%)	**8.71** (8.3%)	**12.22** (13.0%)	**1.96** (50.2%)	**1.03** (30.4%)	**9.39** (21.2%)
CatVTON-FLUX [6]	13.81	5.61	0.844	0.0811	1.26	13.38	16.87	7.34	1.67	15.44
+FitControler	**12.15** (12.0%)	**4.13** (26.4%)	**0.851** (0.8%)	**0.0796** (1.9%)	**0.91** (27.8%)	**9.49** (29.0%)	**15.42** (8.6%)	**6.42** (12.5%)	**1.19** (28.7%)	**10.78** (30.2%)
FitDiT [15]	14.00	4.32	0.841	0.0832	1.26	13.36	15.74	4.30	1.86	17.49
+FitControler	**9.94** (28.9%)	**1.91** (55.7%)	**0.868** (3.1%)	**0.0593** (28.7%)	**0.75** (40.5%)	**7.59** (43.1%)	**12.67** (19.5%)	**2.39** (44.4%)	**1.00** (46.2%)	**8.90** (49.1%)
Leffa [50]	10.56	3.27	0.850	0.0720	1.17	10.39	12.55	4.03	1.31	10.63
+FitControler	**9.83** (6.9%)	**1.90** (41.9%)	**0.864** (1.6%)	**0.0610** (15.3%)	**0.74** (36.8%)	**7.87** (24.3%)	**11.98** (4.5%)	**2.12** (47.4%)	**0.98** (25.2%)	**8.80** (17.2%)

Figure 7: Qualitative results of VTON models with FitControler. Regions showing the most prominent fit variations are highlighted with dashed boxes. Additional examples on other VTON models are provided in the supplementary material.

Hd. Interestingly, improved garment fits also lead to consistently improved performance over other image quality metrics (FID, KID, SSIM, and LPIPS), which suggests that fit is an essential factor in achieving realism, but has been underexplored in previous works. Notably, as discussed in Section 3.3, models such as Leffa [50] rely on masks and dense poses that partially leak the original garment, leading to limited gains from FitControler. In contrast, FitDiT [15] uses square masks and human keypoints to remove fit cues and thus benefits substantially from FitControler, improving all metrics by large margins. In addition, we evaluate performance on a regular-fit-only subset of the short-sleeve data. The results, provided in the supplementary material, follow trends consistent with Table 2. These results confirms that FitControler enables effective fit-aware VTON while enhancing the overall realism via improved fit consistency.

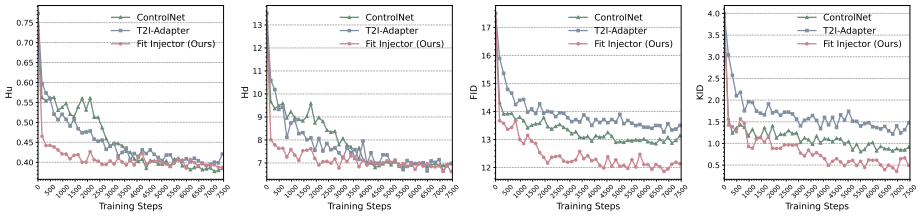

Figure 8: Training curves of different conditional injection approaches. Our approach converges faster with superior image quality.

Table 3: Condition injection. Comparison of fit injector with alternative injection approaches after 7,500 training steps.

Method	FID	KID	Hu	Hd	Trainable Params
ControlNet	13.14	0.91	**0.38**	6.98	361M
T2I-Adapter	13.50	1.48	0.42	6.97	77M
Fit Injector	**12.13**	**0.49**	0.39	**6.63**	97M

Table 4: Training sample size. FitControler trained on 1,000 samples matches full-data performance.

# Samples	FID	KID	Hu	Hd
1000	17.50	1.56	0.44	8.39
2000	17.51	1.42	0.44	**7.21**
3000	**17.46**	**1.31**	**0.44**	8.12
3959 (Full)	17.80	1.52	0.45	8.25

Fig. 1 presents the qualitative results of FitControler on CatVTON [6], and Fig. 7 further shows the results on Leffa and FitDiT. For short-sleeve T-shirts, the generated fits exhibit clear differences in sleeve tightness and torso contour while well preserving garment texture and person identity. For trousers, one can see that *tapered* trousers gradually narrow toward the ankle, and *straight* trousers maintain a uniform leg width. These visualizations intuitively demonstrate that FitControler generates perceptually realistic and well-controlled fits for VTON.

6.4 Ablation Study

Condition Injection. We compare our multi-scale fit injector with alternative controlled image generation approaches including ControlNet [46] and T2I-Adapter [27] on the short-sleeve T-shirts under the paired setting. Table 3 shows that our multi-scale fit injector achieves superior generation quality (FID/KID) while maintaining comparable control ability (Hu/Hd) to ControlNet, but with substantially fewer number of parameters. Fig. 8 illustrates the training curves. Our multi-scale fit injector converges faster and consistently outperforms the baselines in terms of the image quality metrics (FID/KID) throughout the training process. Notably, it can achieve reasonable fit manipulation with only 1,000 training steps.

Training Sample Size. Here we study the impact of training sample size on FitControler by training both the layout generator and fit injector with Leffa [50] on the short-sleeve data of varying sample sizes (1,000, 2,000, 3,000, and 3,959 samples) and test under the unpaired setting. According to Table 4, FitControler achieves good performance even with only 1,000 training samples, implying high sample efficiency.

Table 5: FitControler vs. extra training. FitControler accounts for major improvements on Hu and Hd, while extra finetuning yields only marginal gains.

Method	FitCtrl	Finetune	FID	KID	Hu	Hd
Leffa	✗	✗	18.77	2.59	0.65	10.47
	✗	✓	18.69	2.18	0.63	12.67
	✓	✗	**17.73**	**1.54**	**0.45**	**7.65**
FitDiT	✗	✗	23.94	6.82	0.80	16.89
	✗	✓	19.49	3.55	0.81	14.42
	✓	✗	**18.48**	**1.86**	**0.46**	**7.94**

Table 6: FitControler vs. prompt conditioning. While prompt conditioning improves fit control to some extent, FitControler consistently yields superior fit consistency (Hu/Hd) and overall image quality.

Method	Paired						Unpaired			
	FID↓	KID↓	SSIM↑	LPIPS↓	Hu↓	Hd↓	FID↓	KID↓	Hu↓	Hd↓
Short-sleeve										
IDMVTON	15.99	2.90	0.852	0.0647	0.55	9.03	20.83	3.37	0.67	10.39
+Prompt	15.49	1.98	**0.868**	**0.0581**	0.46	8.38	21.28	3.29	0.56	9.77
+FitControler	**14.69**	**1.92**	0.866	0.0592	**0.45**	**7.54**	**19.11**	**2.19**	**0.51**	**8.81**
Trousers										
IDMVTON	15.11	6.67	0.854	0.0677	1.03	10.53	16.12	5.59	1.58	12.58
+Prompt	13.07	4.14	0.860	**0.0577**	0.89	9.68	15.11	4.96	1.32	11.38
+FitControler	**12.14**	**3.68**	**0.862**	0.0592	**0.81**	**9.32**	**14.43**	**3.95**	**1.13**	**10.29**

FitControler vs. Extra Training. To verify that the observed performance mainly originates from FitControler rather than additional training on Fit4Men, here we conduct an ablation study on short-sleeve shirts. We compare three configurations: (1) *Baseline*: the original VTON model without both FitControler and finetuning; (2) *Backbone finetuning*: finetuning the VTON model on Fit4Men using our garment-agnostic preprocessor, but without FitControler; (3) *FitControler only*: inserting FitControler while freezing the original VTON model. As shown in Table 5, backbone finetuning slightly improves the FID and KID but has negligible effect on fit control. In contrast, incorporating FitControler significantly reduces Hu and Hd, indicating better adherence to the target fit. These results suggest that the performance gains indeed come from FitControler. In contrast, extra training only contributes to minor image quality improvements in this context.

FitControler vs. Prompt Conditioning. To investigate whether fit control can be achieved by text conditioning alone, we construct an IDM-VTON baseline by appending fit labels to the text prompts and fine-tuning the model for 16,000 steps. We find that prompt conditioning alone is ineffective under standard preprocessing due to shape leakage, and only becomes viable when combined with our garment-agnostic representation. As shown in Table 6, although prompt conditioning improves fit consistency over the vanilla IDM-VTON, it consistently underperforms FitControler across both short-sleeve and trouser scenarios. Qualitative comparisons in Fig. 9 further show that prompt-based control yields weaker fit magnitude and less accurate garment boundaries. These results demonstrate that text conditioning alone is insufficient for precise fit control, while explicit spatial modeling in FitControler provides substantially better fit-aware generation.

6.5 Extension to Diverse Poses and Body Shapes

To further evaluate the generalization ability of FitControler, we additionally test it on models with diverse body shapes and poses. As shown in Fig. 10(b) and supplementary Fig. S4, we directly use the original DensePose representation to

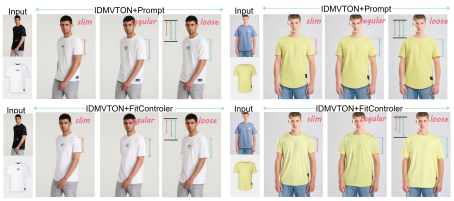

Figure 9: Qualitative comparison of IDMVTON: text control vs. FitControler.

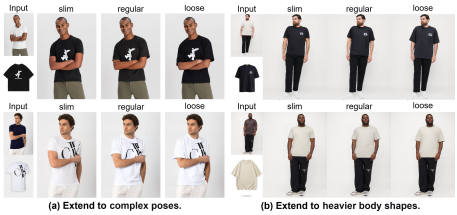

Figure 10: Results of FitControler extended to diverse poses and body shapes.

preserve body-shape information. Despite minor body-shape preservation errors (e.g., slightly thinner appearances under slim-fit settings), our method consistently achieves the desired fit control across different body types. Furthermore, the complex-pose examples in Fig. 10(a) demonstrate that our framework maintains stable fit control under challenging pose variations. These results suggest that FitControler generalizes beyond the controlled training setting, while more diverse body-shape distributions remain an important direction for future study.

7 Conclusion

We introduced FitControler, a novel plug-in module that equips diffusion-based VTON models with precise control over garment fit. It first models garment layouts under specified fit through a fit-aware layout generator conditioned on carefully designed garment-agnostic representations, and then conveys these layout features to VTON models via a multi-scale fit injector, enabling layout-driven generation. In particular, we built a dedicated fit-aware dataset, Fit4Men, and proposed two evaluation metrics specifically tailored to assess fit control. Extensive experiments demonstrate that our approach effectively bridges the gap between high-level fit concepts and realistic try-on image synthesis, providing a practical solution for controllable virtual try-on.

Our current dataset is restricted to menswear, covering only a limited range of garment categories and fit types, and does not sufficiently account for diverse body shapes. These factors may inevitably constrain the diversity and generalization capability of the proposed framework. For future work, we plan to expand the dataset to include womenswear, a broader spectrum of garment categories, richer fit variations, and models with diverse body types, thereby further enhancing the applicability and robustness of fit-aware virtual try-on systems in real-world scenarios.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant No. 62576146.

References

1. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. arXiv preprint arXiv:1801.01401 (2018)
2. Chen, C.Y., Chen, Y.C., Shuai, H.H., Cheng, W.H.: Size does matter: Size-aware virtual try-on via clothing-oriented transformation try-on network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7513–7522 (2023)
3. Chen, M., Chen, X., Zhai, Z., Ju, C., Hong, X., Lan, J., Xiao, S.: Wear-any-way: Manipulable virtual try-on via sparse correspondence alignment. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 124–142. Springer (2024)
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. Choi, Y., Kwak, S., Lee, K., Choi, H., Shin, J.: Improving diffusion models for authentic virtual try-on in the wild. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 206–235. Springer (2024)
6. Chong, Z., Dong, X., Li, H., Zhang, S., Zhang, W., Zhang, X., Zhao, H., Jiang, D., Liang, X.: Catvton: Concatenation is all you need for virtual try-on with diffusion models. arXiv preprint arXiv:2407.15886 (2024)
7. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: International Conference on Machine Learning (2024)
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
9. 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)
10. 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)
11. Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S.: Viton: An image-based virtual try-on network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7543–7552 (2018)
12. 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)
13. Hu, M.K.: Visual pattern recognition by moment invariants. IRE Transactions on Information Theory **8**(2), 179–187 (1962)
14. Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J.: Comparing images using the hausdorff distance. IEEE Transactions on Pattern Analysis and Machine Intelligence **15**(9), 850–863 (1993)
15. Jiang, B., Hu, X., Luo, D., He, Q., Xu, C., Peng, J., Zhang, J., Wang, C., Wu, Y., Fu, Y.: Fitdit: Advancing the authentic garment details for high-fidelity virtual try-on. arXiv preprint arXiv:2411.10499 (2024)

16. Khirodkar, R., Bagautdinov, T., Martinez, J., Zhaoen, S., James, A., Selednik, P., Anderson, S., Saito, S.: Sapiens: Foundation for human vision models. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 206–228 (2024)
17. Kim, J., Gu, G., Park, M., Park, S., Choo, J.: Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8176–8185 (2024)
18. Kim, J., Jin, H., Park, S., Choo, J.: Promptdresser: Improving the quality and controllability of virtual try-on via generative textual prompt and prompt-aware mask. arXiv preprint arXiv:2412.16978 (2024)
19. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
20. Kollnitz, A., Pecorari, M.: Fashion, Performance, and Performativity: The Complex Spaces of Fashion. Bloomsbury Publishing (2021)
21. Labs, B.F.: Flux: Official inference repository for flux.1 models (2024), `https://github.com/black-forest-labs/flux`, accessed: 2024-11-12
22. Lee, S., Gu, G., Park, S., Choi, S., Choo, J.: High-resolution virtual try-on with misalignment and occlusion-handled conditions. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 204–219 (2022)
23. Li, K., Zhang, J., Chang, S.Y., Forsyth, D.: Controlling virtual try-on pipeline through rendering policies. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5866–5875 (2024)
24. Li, Y., Zhou, H., Shang, W., Lin, R., Chen, X., Ni, B.: Anyfit: Controllable virtual try-on for any combination of attire across any scenario. arXiv preprint arXiv:2405.18172 (2024)
25. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
26. 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)
27. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4296–4304 (2024)
28. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
29. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4195–4205 (2023)
30. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
31. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
32. Preechakul, K., Chatthee, N., Wizadwongsa, S., Suwajanakorn, S.: Diffusion autoencoders: Toward a meaningful and decodable representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10619–10629 (2022)

33. Qin, C., Zhang, S., Yu, N., Feng, Y., Yang, X., Zhou, Y., Wang, H., Niebles, J.C., Xiong, C., Savarese, S., et al.: Unicontrol: A unified diffusion model for controllable visual generation in the wild. Advances in Neural Information Processing Systems **36**, 42961–42992 (2023)
34. 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 (2021)
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
36. Wan, S., Chen, J., Pan, Y., Yao, T., Mei, T.: Incorporating visual correspondence into diffusion model for virtual try-on. In: International Conference on Learning Representations (2025)
37. 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)
38. 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)
39. Xu, Y., He, Z., Shan, S., Chen, X.: Ctrlora: An extensible and efficient framework for controllable image generation. In: International Conference on Learning Representations (2025)
40. Xu, Y., Gu, T., Chen, W., Chen, A.: Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 8996–9004 (2025)
41. Yan, K., Gao, T., Zhang, H., Xie, C.: Linking garment with person via semantically associated landmarks for virtual try-on. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17194–17204 (2023)
42. Yang, X., Ding, C., Hong, Z., Huang, J., Tao, J., Xu, X.: Texture-preserving diffusion models for high-fidelity virtual try-on. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7017–7026 (2024)
43. Yang, Z., Zeng, A., Yuan, C., Li, Y.: Effective whole-body pose estimation with two-stages distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4210–4220 (2023)
44. Yu, R., Wang, X., Xie, X.: Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10511–10520 (2019)
45. Yu, W.W.m., Hunter, L.: Clothing appearance and fit: Science and technology. Woodhead publishing (2004)
46. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
47. 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/CVF Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018)
48. Zhang, X., Song, D., Zhan, P., Chang, T., Zeng, J., Chen, Q., Luo, W., Liu, A.A.: Boow-vton: Boosting in-the-wild virtual try-on via mask-free pseudo data training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26399–26408 (2025)

49. Zhao, S., Chen, D., Chen, Y.C., Bao, J., Hao, S., Yuan, L., Wong, K.Y.K.: Uni-controlnet: All-in-one control to text-to-image diffusion models. Advances in Neural Information Processing Systems **36**, 11127–11150 (2023)
50. Zhou, Z., Liu, S., Han, X., Liu, H., Ng, K.W., Xie, T., Cong, Y., Li, H., Xu, M., Pérez-Rúa, J.M., et al.: Learning flow fields in attention for controllable person image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2491–2501 (2025)
51. Zhu, L., Li, Y., Liu, N., Peng, H., Yang, D., Kemelmacher-Shlizerman, I.: M&m vto: Multi-garment virtual try-on and editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1346–1356 (2024)
52. 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)