Rethinking Garment Conditioning in Diffusion-based Virtual Try-On: Decouple, Don't Denoise

Kihyun Na[1], Jinyoung Choi[2], and Injung Kim[2] ✉

[1] Advanced AI Talent Education and Research Group for Industrial Innovation, Handong Global University, Pohang 37554, Republic of Korea
khna@handong.edu

[2] Department of Artificial Intelligence, Computer and Electrical Engineering, Handong Global University, Pohang 37554, Republic of Korea
jinyoung@handong.ac.kr, ijkim@handong.edu

Abstract. Virtual Try-On (VTON) synthesizes realistic images of a person wearing a target garment, with broad applications in e-commerce and fashion. Diffusion-based dual-UNet methods achieve strong results but double the parameters by dedicating a separate network to garment conditioning. Spatial concatenation offers a simpler single-network alternative, yet both UNet- and DiT-based instantiations report that full fine-tuning is ineffective, and the community has settled for attention-only training. We ask: why does full fine-tuning fail, and can this be resolved? Through what is, to our knowledge, the first visualization study of dual-UNet reference network behavior, we identify a unifying insight: garment conditioning must be decoupled from the denoising process. Spatial concatenation violates this by embedding the garment within the denoising target, causing three conflicts: guidance leakage, gradient competition, and train-test discrepancy. We derive three design principles to restore this decoupling and implement them as a pure recipe atop a standard architecture with no modification. The resulting model, DeCo-VTON (860M params), achieves single-network state of the art, matching the dual-UNet state of the art at half the cost while being preferred in human evaluation.

Keywords: Virtual Try-On · Diffusion Models

1 Introduction

Image-based virtual try-on (VTON) aims to synthesize a realistic image of a person wearing a target garment, given a person image and a garment image as input. With broad applications in fashion e-commerce—from enhancing customer experience [6, 19, 56] to reducing return rates [43]—VTON has attracted growing attention as a conditional image generation task. While early approaches relied on Generative Adversarial Networks [3, 45, 47, 48], diffusion-based methods have become the dominant paradigm owing to their superior generation quality [4, 5, 12, 23, 34, 49, 57].

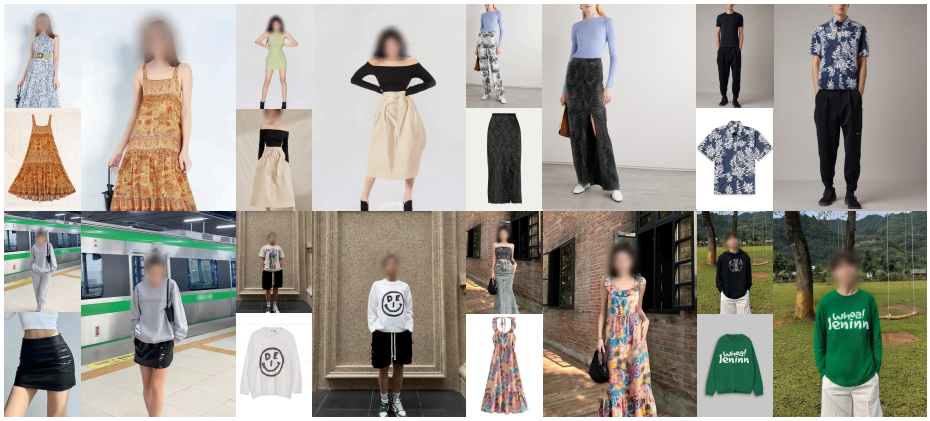

Fig. 1: Virtual Try-on images generated by our DeCo-VTON on the Public datasets and In-the-Wild images. Please zoom in for more details.

Among diffusion-based approaches, dual-UNet architectures [4,49,57] achieve strong garment fidelity by dedicating a separate reference UNet to garment conditioning. However, this design doubles the parameter count and computational cost. Single-network alternatives have thus attracted interest for their efficiency, with *spatial concatenation*—combining garment and person latents into a single denoising input—emerging as the most straightforward approach [5]. Yet a puzzling pattern has emerged: both CatVTON [5] (860M UNet) and Voost [27] (11.9B DiT) report that full fine-tuning is ineffective under spatial concatenation, and both settle for attention-only training. Despite differing in architecture and scale by an order of magnitude, they reach the same conclusion. **Why does full fine-tuning fail, and can it be resolved?**

In this paper, we answer this question. We begin by analyzing the noise prediction behavior of dual-UNet reference networks—to our knowledge, the first such visualization study—and identify a unifying insight behind their effectiveness: *the garment conditioning pathway must be decoupled from the denoising process.* Spatial concatenation violates this insight by embedding the garment latent within the denoising target. This loss of decoupling gives rise to three functional conflicts: (1) garment information leaks into the unconditional branch, causing classifier-free guidance to suppress rather than enhance garment details; (2) the training objective includes garment reconstruction, creating gradient competition between reconstruction and conditioning; and (3) at inference time, the garment latent is updated by model predictions rather than the forward diffusion schedule, introducing a train-test discrepancy that accumulates over denoising steps.

We derive three design principles that restore this decoupling within the spatial-concatenation framework: Garment-Free Guidance, Decoupled Loss, and Clean Latent Anchoring, and implement them as a training and inference recipe on a standard architecture (SD1.5 inpainting UNet) with no architectural mod-

ification. The resulting model, DeCo-VTON, is the first to unlock effective full fine-tuning under spatial concatenation. With only 860M parameters, DeCo-VTON achieves state-of-the-art results among single-network methods, reaching comparable quality to the dual-UNet state of the art [57] at half the parameters, 2.1× faster inference, and 42% lower VRAM, while being preferred in human evaluation. Fig. 1 shows try-on results of DeCo-VTON on the public benchmarks and on in-the-wild images.

Our contributions are as follows: (1) **Analysis:** We present the first visualization study of dual-UNet reference network behavior, revealing that successful garment conditioning requires decoupling from the denoising process. (2) **Diagnosis & Principles:** We identify three functional conflicts under spatial concatenation and derive three design principles that resolve each and unlock full fine-tuning. (3) **Results:** Without architectural modification, DeCo-VTON achieves single-network state of the art and matches the dual-UNet state of the art at half the cost.

2 Related Work

Early VTON models adopted GAN-based [11, 22] pipelines, typically comprising a warping module for geometric alignment followed by a conditional generator for blending. VITON [15] established this framework using Thin-Plate Spline (TPS) [2] transformations, and subsequent methods [3, 8, 45, 47, 48, 50] refined geometric matching through contextual cues and human parsing. While effective, these approaches depend on external modules such as pose estimators, and often produce artifacts under challenging poses or complex textures.

The advent of latent diffusion models (LDMs) [41], which operate in a compact latent space learned by a Variational Auto-Encoder (VAE) [25], brought stronger generative priors and has since become the dominant paradigm for VTON [12, 23, 34, 51]. Diffusion-based VTON methods mainly differ in how they inject garment information into the denoising backbone, typically a UNet [42] derived from Stable Diffusion [41] or SDXL [39]. Dual-UNet architectures introduce a dedicated reference UNet that processes the garment image and provides multi-scale features to the denoising UNet. TryOnDiffusion [58] pioneered this design using two parallel UNets with implicit warping through cross-attention. OOTDiffusion [49] adopts the outfitting fusion approach, where the reference UNet features are injected via self-attention layers of the denoising UNet. IDM-VTON [4] combines high-level garment descriptors from IP-Adapter [52] with low-level feature maps from the reference UNet. Most recently, Leffa [57] introduces a flow-field attention loss and additional densepose conditioning [13], achieving the current dual-UNet state of the art. Despite strong fidelity, all dual-UNet methods incur substantially higher computational cost due to the additional network.

Single-network architectures remove the reference UNet to reduce overhead. LaDI-VTON [34] encodes garments as pseudo-word embeddings via textual inversion [9]. DCI-VTON [12] adds an explicitly warped garment [14] with the

masked person. CatVTON [5] simplifies this further through spatial concatenation of the raw garment and masked person latents, relying solely on internal self-attention without requiring additional encoders such as CLIP [40] or DINOv2 [36]. More recently, Voost [27] applies the same spatial-concatenation principle to a DiT backbone at 11.9B parameters. Notably, both CatVTON and Voost report that full fine-tuning is ineffective under spatial concatenation and instead adopt attention-only training—an observation we analyze in depth in Section 3.

From a broader perspective, conditioning in diffusion models ranges from cross-attention tokens to dedicated parallel networks [53] to direct input-level concatenation. In the first two families the condition stays structurally separate from the denoising target—in per-modality streams of multimodal transformers [7] or a dedicated branch of inpainting models [20]—and is thus never subject to the reconstruction loss. Spatial concatenation instead embeds the condition within the target, making the garment at once a conditioning signal and a reconstruction target; this entanglement is the implication this paper systematically studies.

3 Analyzing Garment Conditioning

3.1 How Dual UNets Condition on Garments

In dual-UNet VTON architectures, a dedicated reference UNet processes the garment image and provides conditioning features to the main denoising UNet. Although these models share a common high-level design, their reference UNet strategies differ substantially: IDM-VTON [4] freezes the pretrained weights, OOTDiffusion [49] fine-tunes but fixes the timestep at t=0, and Leffa [57] fine-tunes while feeding the same timestep t used by the denoising UNet. These design choices lead to significant performance differences, yet no prior work has systematically analyzed why.

We briefly establish notation. In latent diffusion models, the forward process produces a noisy latent $\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}$ at timestep t, where $\bar{\alpha}_t$ is the cumulative noise schedule and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. A noise predictor $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t)$ is trained to estimate $\boldsymbol{\epsilon}$ from the noisy input, and we denote by $\mathbf{z}_0^g = \mathcal{E}(I^g)$ the clean garment latent obtained from the VAE encoder $\mathcal{E}$.

To understand the reference UNet's behavior, we visualize its noise prediction output $\boldsymbol{\epsilon}_t^g$ at different timesteps by passing it through the VAE decoder, as shown in Fig. 2. In standard diffusion models, the noise predictor $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t)$ recovers different information depending on the timestep: primarily low-frequency, global structure at high t and high-frequency details at low t [17, 21, 29, 37]. Since the reference UNet shares the same architecture, visualizing its output across timesteps reveals how it provides conditioning information to the denoising process. To the best of our knowledge, this is the first analysis of reference UNet behavior through noise prediction visualization.

IDM-VTON uses a frozen pretrained SDXL UNet as the reference network. It receives the clean garment latent $\mathbf{z}_0^g$ as input together with the denoising

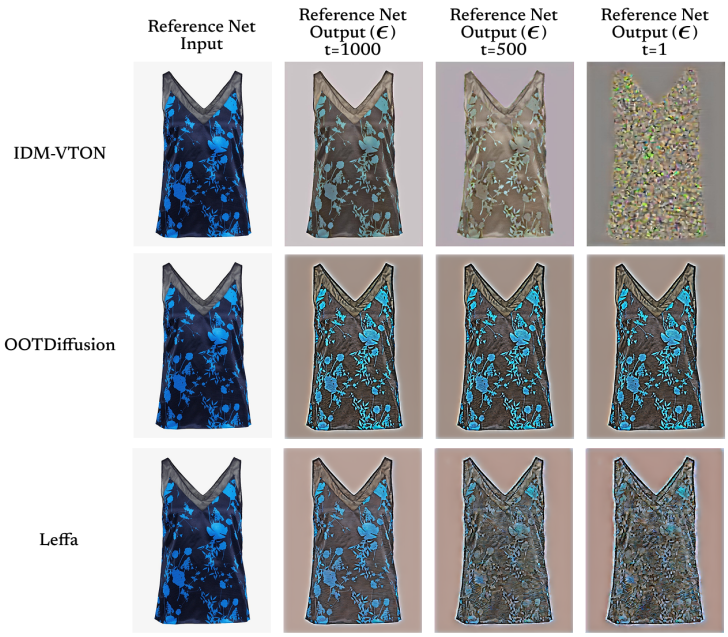

Fig. 2: Visualization of the noise prediction $\boldsymbol{\epsilon}_t^g$ from the reference UNets at key timesteps. **Top:** IDM-VTON (frozen) produces strong garment features only at high t, collapsing near t=1 due to the mismatch between its clean input and the pretrained noise-level expectation. **Middle:** OOTDiffusion produces strong but timestep-invariant features due to its fixed t=0 design. **Bottom:** Leffa (fine-tuned with varying t) produces timestep-dependent features that transition from coarse structure to fine detail, reflecting the garment consistently across all stages.

timestep t. However, the pretrained weights expect noisy input whose statistics match timestep t: at high t, the network assumes heavily corrupted input, creating a severe distributional discrepancy with the actual clean $\mathbf{z}_0^g$—a mismatch independently observed in diffusion feature extraction [44]. As a result, the extracted features become unreliable at high t (Fig. 2, top). At low t, the discrepancy diminishes but the output exhibits low variance, limiting garment conditioning quality across the full trajectory.

OOTDiffusion addresses the input mismatch by fine-tuning the reference UNet for the VTON task, enabling it to extract meaningful garment features from clean input. However, its timestep is fixed at t=0 during both training and inference (Fig. 2, middle). While this produces strong garment features, the output remains identical regardless of the current denoising stage—the network cannot modulate its features to match the coarse-to-fine progression of the denoising process.

Leffa fine-tunes the reference UNet while feeding it the same timestep t used by the denoising UNet. This combination allows the reference UNet to learn stage-

adaptive garment features: at high t, it provides coarse structural information; at low t, it shifts to fine-grained texture details (Fig. 2, bottom). This timestep-aligned conditioning is the key factor behind Leffa's superior performance among dual-UNet methods.

Comparing these three designs reveals a unifying insight: **successful garment conditioning requires the conditioning pathway to be decoupled from the denoising process.** In dual-UNet architectures, this decoupling is structural: the reference UNet operates on clean garment input in a separate forward pass and is not subject to the denoising objective. IDM-VTON partially benefits from this separation but fails to fully exploit it due to its frozen weights. Leffa achieves the strongest conditioning because it combines the decoupled pathway with timestep-aligned feature generation. As we show next, spatial concatenation fundamentally eliminates this structural decoupling.

3.2 The Functional Mismatch in Spatial Concatenation

Spatial concatenation combines the garment and person latents into a single input, applies noise to the entire composite, and denoises it as a whole. This design eliminates the need for a separate reference network, offering significant parameter efficiency. However, the garment latent becomes part of the denoising target—the structural decoupling identified in Sec. 3.1 is removed.

This loss of decoupling gives rise to three functional conflicts:

Conflict 1: Garment leakage into classifier-free guidance. Classifier-free guidance (CFG) [18] amplifies the difference between conditional and unconditional predictions to strengthen the conditioning signal. Under spatial concatenation, the unconditional prediction is generated with the garment latent still present in the input, since it is part of the denoising target rather than an external condition. The difference between the two branches thus arises from factors other than the garment, and increasing the guidance scale ω suppresses rather than enhances garment details.

Conflict 2: Gradient competition between reconstruction and conditioning. The training objective is computed over the entire denoising target, including the garment region. The network receives competing gradients: one directing it to accurately reconstruct the garment, and another directing it to transfer garment information to the person region. These objectives are not aligned—features optimized for pixel-level garment reconstruction are not necessarily optimal for conditioning person generation. In dual-UNet architectures, the reference UNet carries no reconstruction loss and serves purely as a feature extractor.

Conflict 3: Train-test discrepancy in the garment latent. During training, forward diffusion provides the garment region with a correctly scheduled noisy latent at each timestep, ensuring that the noise level matches the network's expectation. At inference, however, the garment latent is updated by the model's own predictions at each step, deviating from the forward diffusion schedule. This discrepancy accumulates over steps, and the timestep-aligned conditioning that

made Leffa successful (Sec. 3.1) breaks down—the garment region's noise level no longer matches what the network expects at timestep t.

These conflicts manifest at two observable levels. Conflicts 2 and 3 degrade the conditional branch itself: even without CFG (ω=1.0), garment details appear blurred because the network's garment encoding capacity is compromised. Conflict 1 causes further degradation when CFG is applied: increasing ω amplifies garment suppression rather than enhancement. We provide quantitative evidence for both effects in Sec. 5.5 and 5.6.

Cross-architecture evidence supports this analysis. CatVTON [5] (860M UNet) reports that full fine-tuning yields no improvement over attention-only training. Voost [27] (11.9B DiT) reports that full fine-tuning (FID 6.351) underperforms attention-only training (FID 5.269) with less than a quarter of the trainable parameters. The same conclusion emerging from fundamentally different architectures and scales strongly suggests that the problem lies in spatial concatenation itself, not in any particular backbone.

We argue that the ineffectiveness of full fine-tuning observed in previous studies stems from these unresolved conflicts rather than from an intrinsic limitation of full fine-tuning itself. If the three conflicts can be explicitly resolved within the spatial-concatenation framework, full fine-tuning should unlock stronger garment conditioning by allowing the entire network to adapt to the try-on task. We derive three design principles—each resolving one conflict—and present their implementations in the following section.

4 Decoupled Conditioning for Virtual Try-On

We instantiate three design principles—each resolving one conflict from Sec. 3.2—as a training and inference recipe applied to CatVTON [5], deliberately keeping the architecture identical: same backbone (SD1.5 inpainting UNet, 860M parameters), same input formulation, same optimizer and learning rate.

4.1 Overview

CatVTON conditions the denoising process through spatial concatenation of garment and person latents. Given a masked person image I^p and a garment image I^g, the VAE encoder produces latents $\mathbf{z}_0^p = \mathcal{E}(I^p)$ and $\mathbf{z}_0^g = \mathcal{E}(I^g)$. These are spatially concatenated along the height dimension, and the input to the denoising UNet at timestep t is:

$$\mathbf{X}_t = \mathrm{Concat}_{\mathrm{ch}}(\mathbf{z}_t^p \oplus \mathbf{z}_t^g,\ \mathbf{M} \oplus \mathbf{0},\ \mathbf{z}_0^p \oplus \mathbf{z}_0^g)\,, \tag{1}$$

where $\oplus$ denotes spatial concatenation along the height dimension, $\mathrm{Concat}_{\mathrm{ch}}$ denotes channel-wise concatenation, and $\mathbf{M}$ is the inpainting mask. The standard training objective computes the noise prediction loss over the entire spatially concatenated output:

$$\mathcal{L}_{\mathrm{standard}} = \mathbb{E}_{\mathbf{z}_0,\boldsymbol{\epsilon},t}\Big[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{X}_t, t)\|^2\Big]\,, \tag{2}$$

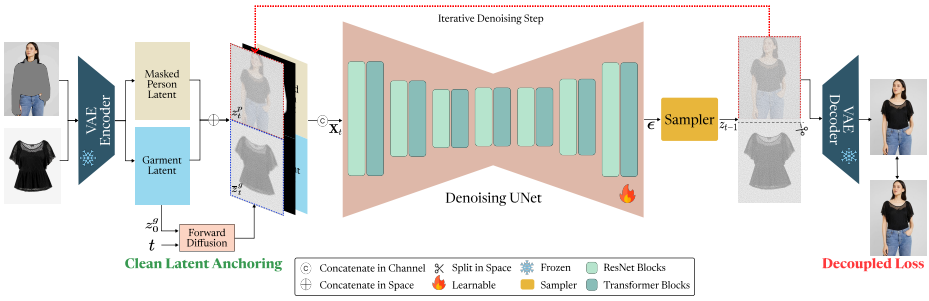

Fig. 3: Overview of DeCo-VTON. The masked person image and the garment image are encoded into latent representations using a VAE and spatially concatenated to form the denoising input. The Decoupled Loss supervises only the person region, allowing the garment region to serve purely as a conditioning signal. At each inference step, Clean Latent Anchoring replaces the garment latent with the forward-diffused ground truth, ensuring timestep-aligned conditioning throughout the denoising trajectory. The final latent is decoded to produce the try-on output.

where $\boldsymbol{\epsilon} = \boldsymbol{\epsilon}^p \oplus \boldsymbol{\epsilon}^g$ includes noise from both regions. As analyzed in Sec. 3.2, this formulation—combined with the standard CFG and inference procedure—gives rise to three conflicts that prevent full fine-tuning from succeeding.

DeCo-VTON applies full fine-tuning to the entire UNet and introduces a training and inference recipe comprising three principles (Fig. 3): Garment-Free Guidance (Sec. 4.2), Decoupled Loss (Sec. 4.3), and Clean Latent Anchoring (Sec. 4.4).

4.2 Garment-Free Guidance

Conflict 1 (Sec. 3.2) arises because the unconditional branch in standard CFG retains garment information. In dual-UNet methods, disabling the reference network naturally yields a garment-free unconditional prediction. Spatial concatenation requires an explicit mechanism to achieve the same effect.

In CatVTON's CFG, the unconditional input removes the clean garment latent $\mathbf{z}_0^g$ from the conditioning channels but retains the noisy garment latent $\mathbf{z}_t^g$:

$$\mathbf{X}_t^{\text{uncond}} = \text{Concat}_{\text{ch}}(\mathbf{z}_t^p \oplus \mathbf{z}_t^g,\ \mathbf{M} \oplus \mathbf{0},\ \mathbf{z}_0^p \oplus \mathbf{0})\,. \tag{3}$$

The retained $\mathbf{z}_t^g$ leaks garment information into the unconditional prediction, causing CFG to suppress garment details as the guidance scale ω increases.

We construct a strictly garment-free unconditional branch by removing all garment-related latents:

$$\mathbf{X}_t^{\text{gf}} = \text{Concat}_{\text{ch}}(\mathbf{z}_t^p \oplus \mathbf{0},\ \mathbf{M} \oplus \mathbf{0},\ \mathbf{z}_0^p \oplus \mathbf{0})\,. \tag{4}$$

The guided prediction then becomes:

$$\hat{\boldsymbol{\epsilon}}_t = \boldsymbol{\epsilon}_\theta(\mathbf{X}_t^{\text{gf}}, t) + \omega\Big(\boldsymbol{\epsilon}_\theta(\mathbf{X}_t, t) - \boldsymbol{\epsilon}_\theta(\mathbf{X}_t^{\text{gf}}, t)\Big) \tag{5}$$

By making the unconditional branch entirely garment-free, the guidance signal, *i.e.* the difference between the two predictions, now correctly isolates the garment's contribution; the conditional branch retains full garment information, so the guidance amplifies rather than suppresses garment details. We term this scheme Garment-Free Guidance (GFG), where *garment-free* refers specifically to the unconditional branch, analogous to how disabling the reference network in a dual-UNet model yields an unconditional prediction that serves as a clean person prior without any garment information. During training, we randomly drop the garment condition with probability 10% to enable the model to learn the garment-free prediction.

4.3 Decoupled Loss

Conflict 2 (Sec. 3.2) arises because the training objective in Eq. 2 includes the garment region, creating gradient competition between garment reconstruction and person generation. In dual-UNet methods, the reference UNet carries no reconstruction loss and serves purely as a feature extractor. To replicate this separation, we exclude the garment region from the training objective.

Since the person and garment latents are spatially concatenated along the height dimension (person on top, garment on bottom), the full UNet output naturally decomposes as $\boldsymbol{\epsilon}_\theta = \boldsymbol{\epsilon}^p_\theta \oplus \boldsymbol{\epsilon}^g_\theta$, where $\boldsymbol{\epsilon}^p_\theta$ denotes the person-region portion. We compute the loss only on the person region:

$$\mathcal{L}_{\text{decoupled}} = \mathbb{E}_{\mathbf{z}_0,\boldsymbol{\epsilon},t}\left[\left\|\bar{\boldsymbol{\epsilon}}^p_{\text{dream},t} - \boldsymbol{\epsilon}^p_\theta(\tilde{\mathbf{X}}_t, t)\right\|^2\right], \tag{6}$$

where we adopt the DREAM [55] objective with its target rectification and input rectification applied exclusively to the person region, keeping the garment region fixed. The rectified input $\tilde{\mathbf{X}}_t$ is constructed by substituting the DREAM-corrected person latent into Eq. 1 while leaving the garment latent unchanged. No mask generation or pixel-wise weighting is required—the spatial layout directly enables the decomposition. Full details are provided in the supplementary material.

4.4 Clean Latent Anchoring

Conflict 3 (Sec. 3.2) arises because the garment latent at inference is updated by the model's own predictions, deviating from the forward diffusion schedule and breaking the timestep-aligned conditioning that made Leffa successful in the dual-UNet setting (Sec. 3.1).

We distinguish $\bar{\mathbf{z}}^g_t$ (computed from $\mathbf{z}^g_0$ via forward diffusion) from $\mathbf{z}^g_t$ (updated by the model's prediction chain during standard inference). Since the clean garment latent $\mathbf{z}^g_0$ is available as input, we compute the ground-truth noisy latent at any timestep via forward diffusion:

$$\bar{\mathbf{z}}^g_t = \sqrt{\bar{\alpha}_t}\,\mathbf{z}^g_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}_{\text{init}}, \quad \boldsymbol{\epsilon}_{\text{init}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{7}$$

At each denoising step, we replace $\mathbf{z}_t^g$ with $\bar{\mathbf{z}}_t^g$ in Eq. 1. A fixed noise sample $\boldsymbol{\epsilon}_{\text{init}}$ is used throughout the denoising process to maintain consistency. This ensures two properties: (1) no error accumulation, since the garment latent is always derived directly from $\mathbf{z}_0^g$ rather than from the prediction chain; and (2) exact timestep alignment, since the noise level in the garment region matches what the network expects at timestep t.

CLA is an inference-only modification. During training, forward diffusion already provides the garment region with correctly scheduled noisy latents, so the training procedure requires no change. This alignment between training and inference distributions is precisely what CLA restores. We note that conditioning the sampling process on a known signal has been explored in diffusion-based inpainting [32] and noising-denoising editing [33]. CLA adopts a similar known-signal anchoring during sampling but serves a distinct purpose: rather than inpainting unknown pixels, it maintains timestep-aligned garment conditioning within the spatial-concatenation framework (Sec. 3.1), preventing error accumulation in the conditioning (not target) region.

5 Experiments

5.1 Setup

Datasets. We evaluate on two public benchmarks: VITON-HD [3] (11,647 training / 2,032 testing upper-body pairs) and DressCode [35] (48,392 training / 5,400 testing full-body pairs across tops, bottoms, and dresses). For DressCode, garment-agnostic masks are generated using DensePose [13] and the SCHP parser [10, 28, 30].

Implementation. We adopt the inpainting variant of Stable Diffusion 1.5 [41] as the backbone—identical to CatVTON [5]. Training is conducted at 512×384 resolution using AdamW [31] with batch size 128, learning rate 1×10^{-5}, and garment dropout probability of 0.1. We train for 16K steps on VITON-HD and 32K steps on DressCode on 2 NVIDIA H200 GPUs (approximately 10 and 20 hours, respectively). The guidance scale is set to ω=2.5 following [5, 49].

Metrics. We report FID [16, 38], KID [1], SSIM [46], and LPIPS [54] for paired evaluation, and FID and KID for unpaired evaluation.

5.2 Quantitative Comparison

Tables 1 and 2 compare DeCo-VTON against UNet-based methods on VITON-HD and DressCode. On VITON-HD, DeCo-VTON achieves the best FID, KID, and LPIPS in the paired setting, trailing Leffa only in SSIM. On DressCode, DeCo-VTON matches Leffa in LPIPS and yields a lower KID, while achieving the best unpaired FID. Notably, DeCo-VTON uses a single 860M UNet, whereas Leffa requires a dual-UNet architecture with 1.8B parameters. These results demonstrate that resolving the functional mismatch through our conditioning recipe is sufficient to close the single–dual network performance gap without

Table 1: Quantitative comparison on VITON-HD. **Best** in bold, second-best underlined. All methods are UNet/CNN-based; for DiT-based comparisons, see Appendix B.

Method	Paired				Unpaired	
	FID↓	KID↓	SSIM↑	LPIPS↓	FID↓	KID↓
HR-VTON [26]	10.88	4.480	0.876	0.097	13.06	4.720
GP-VTON [47]	8.726	3.944	0.870	0.059	11.84	4.310
DCI-VTON [12]	9.408	4.547	0.862	0.061	12.53	5.251
LaDI-VTON [34]	6.660	1.080	0.876	0.091	9.410	1.600
StableVITON [23]	6.439	0.942	0.854	0.091	11.05	3.914
IDM-VTON [4]	5.762	0.732	0.850	0.060	14.65	8.754
PromptDresser [24]	8.540	0.670	0.869	0.112	–	–
OOTDiffusion [49]	9.305	4.086	0.819	0.088	12.41	4.689
CatVTON [5]	5.425	0.411	0.870	0.057	9.015	1.091
Leffa [57]	4.540	0.050	**0.899**	0.048	8.520	**0.320**
DeCo-VTON (Ours)	**4.438**	**0.010**	0.880	**0.047**	**8.266**	0.517

Table 2: Quantitative comparison on DressCode. Notation follows Table 1.

Method	Paired				Unpaired	
	FID↓	KID↓	SSIM↑	LPIPS↓	FID↓	KID↓
GP-VTON [47]	9.927	4.610	0.771	0.180	12.79	6.627
LaDI-VTON [34]	9.555	4.683	0.766	0.237	10.68	5.787
IDM-VTON [4]	6.821	2.924	0.880	0.056	9.546	4.320
OOTDiffusion [49]	4.610	0.955	0.885	0.053	12.57	6.627
CatVTON [5]	3.992	0.818	0.892	0.046	6.137	1.403
Leffa [57]	**2.060**	0.070	**0.924**	**0.031**	4.480	**0.620**
DeCo-VTON (Ours)	2.175	**0.062**	0.914	**0.031**	**4.310**	0.628

architectural changes. Results at higher resolution (1024×768) are reported in Appendix B, where DeCo-VTON likewise attains the best paired FID on both benchmarks, confirming that the recipe remains effective at high resolution.

5.3 Efficiency Analysis

DeCo-VTON shares the identical backbone with CatVTON, so its computational cost is unchanged: 860M parameters, 974 GFLOPs, 1.3s latency, and 2.26 GB peak VRAM (Table 3). Compared to Leffa, DeCo-VTON achieves comparable quality at 2.1× faster inference and 42% lower memory, confirming that a properly designed conditioning recipe can match dual-UNet performance without the associated overhead. Measurement details are provided in the supplementary material.

Table 3: Efficiency comparison at 512×384 resolution. Latency and memory are measured on a single H200 GPU with FP16.

Method	Params (M)	GFLOPs	Latency (s)	VRAM (GB)
OOTDiffusion [49]	2229.73	1225.16	1.5	5.93
IDM-VTON [4]	7003.26	2679.45	6.6	14.62
CatVTON [5]	859.54	973.99	1.3	2.26
Leffa [57]	1802.72	1012.03	2.7	3.91
DeCo-VTON (Ours)	**859.54**	**973.99**	**1.3**	**2.26**

Table 4: Progressive ablation on VITON-HD. GFG: Garment-Free Guidance, DL: Decoupled Loss, CLA: Clean Latent Anchoring. CatVTON results are from the original paper [5]. Each row adds one principle to the previous configuration.

Method	Paired				Unpaired	
	FID↓	KID↓	SSIM↑	LPIPS↓	FID↓	KID↓
CatVTON (Attn-only)	5.425	0.411	0.870	0.057	9.015	1.091
CatVTON (Full FT)	5.250	0.402	0.869	0.055	8.813	0.956
+ GFG	4.538	0.074	0.880	0.048	8.374	0.610
+ DL	4.517	0.068	0.880	0.048	8.338	0.596
+ CLA (= DeCo-VTON)	**4.438**	**0.010**	**0.880**	**0.047**	**8.266**	**0.517**

5.4 Qualitative Comparison

Fig. 4 shows qualitative results on both datasets. DeCo-VTON preserves fine garment details—logos, textures, and complex patterns—more faithfully than CatVTON, and produces results visually comparable to Leffa.

5.5 Ablation Study

Table 4 validates each principle through progressive addition. Two observations directly support the analysis in Sec. 3.2.

First, CatVTON with full fine-tuning (row 2) yields only marginal improvement over attention-only training (row 1), consistent with the functional mismatch preventing full FT from reaching its potential.

Second, resolving Conflict 1 via GFG (row 3) yields the largest single gain (FID: 5.250→4.538, KID: 0.402→0.074), confirming that garment leakage into the unconditional branch is the most severe bottleneck under spatial concatenation. Decoupled Loss (row 4) and CLA (row 5) provide further gains by eliminating gradient competition and restoring timestep alignment, respectively, with CLA yielding a notable KID reduction (0.068→0.010).

We further isolate DL's individual contribution with a leave-one-out comparison at both resolutions in Appendix D, where its effect is robust to the ablation ordering at 512×384 and becomes more pronounced at 1024×768.

Fig. 4: Qualitative comparison on VITON-HD and DressCode. DeCo-VTON preserves garment details (logos, text, patterns) more faithfully than CatVTON and produces results comparable to the dual-UNet model Leffa, despite using a single UNet with half the parameters.

5.6 Diagnostic Analysis: CFG Scale

Fig. 5 provides the diagnostic evidence referenced in Sec. 3.2. At ω=1.0 (no guidance), CatVTON already produces blurred garment details, confirming that Conflicts 2 and 3 degrade the conditional branch itself—the network's garment encoding capacity is compromised even before CFG is applied. Increasing ω further degrades CatVTON's output with artifacts and over-sharpening, confirming

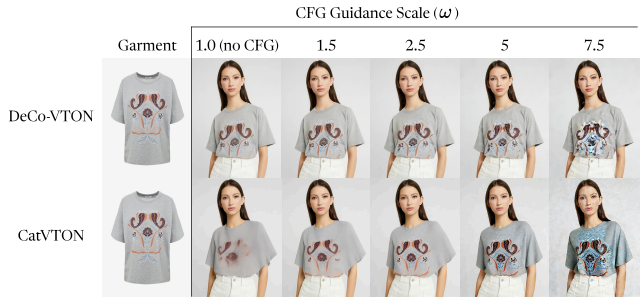

Fig. 5: Effect of guidance scale ω. **Top:** DeCo-VTON preserves garment details even at ω=1.0 and remains stable across scales. **Bottom:** CatVTON produces blurred garments at ω=1.0 (Conflicts 2&3) and introduces artifacts at higher ω (Conflict 1).

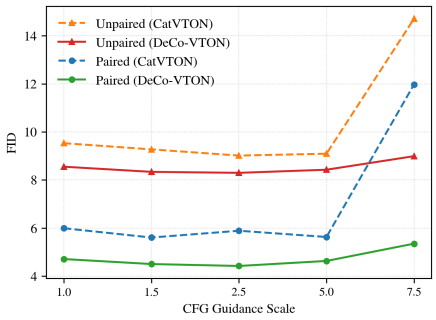

Fig. 6: FID vs. guidance scale ω on VITON-HD. CatVTON degrades sharply at higher ω; DeCo-VTON remains stable.

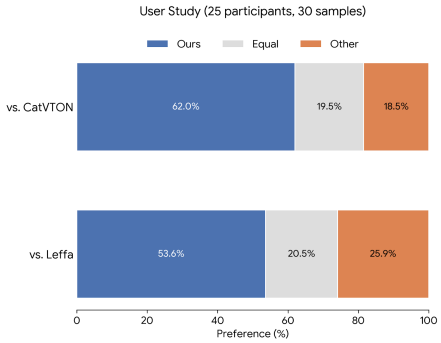

Fig. 7: User study (25 participants, 30 samples; both p<0.001). DeCo-VTON is preferred over CatVTON at approximately 3.5:1 and over Leffa at approximately 2:1.

Conflict 1: the unconditional branch retains garment information, causing guidance to suppress rather than enhance garment details.

In contrast, DeCo-VTON preserves readable logos at ω=1.0 and remains visually stable across scales, validating that the three conflicts have been resolved. Fig. 6 quantitatively confirms this trend: CatVTON's FID degrades sharply beyond ω=2.5, whereas DeCo-VTON remains stable across all tested scales.

5.7 User Study

Fig. 7 reports a randomized blind user study. DeCo-VTON is preferred over CatVTON at approximately 3.5:1 and over Leffa at approximately 2:1, confirming that the quantitative gains translate to perceptually meaningful improvements.

6 Conclusion

This paper asked why full fine-tuning fails in spatial-concatenation-based virtual try-on and showed that the answer lies in a single insight: *garment conditioning must be decoupled from the denoising process.* By analyzing dual-UNet reference networks through noise prediction visualization, we identified that spatial concatenation violates this insight by embedding the garment within the denoising target, giving rise to three functional conflicts—guidance leakage, gradient competition, and train-test discrepancy. Three design principles (Garment-Free Guidance, Decoupled Loss, and Clean Latent Anchoring) restore the necessary decoupling, thereby unlocking effective full fine-tuning for the first time under spatial concatenation.

The resulting model, DeCo-VTON, achieves state-of-the-art results among single-network methods and matches the dual-UNet state of the art at half the parameters, with no architectural modification. This demonstrates that a properly designed conditioning recipe alone can unlock the potential of an existing backbone—the architecture was never the bottleneck; the unresolved conflict between conditioning and denoising was.

Limitations and future work. Our principles are validated on UNet-based spatial concatenation with classifier-free guidance. Flow-matching architectures such as Voost [27] exhibit the same full fine-tuning failure (Sec. 3.2), suggesting the mismatch is architecture-agnostic.

Among our three principles, only Garment-Free Guidance is tied to the CFG mechanism; Decoupled Loss and Clean Latent Anchoring instead address gradient competition and the train-test discrepancy, which arise from spatial concatenation itself and are thus independent of the guidance scheme, so we expect them to transfer to DiT and flow-matching backbones with only minor adaptation. Adapting GFG to non-CFG architectures would instead require an alternative mechanism that prevents garment information from influencing the baseline prediction without relying on a conditional-unconditional split.

More broadly, the conditioning–denoising conflict we identify is not specific to virtual try-on: any spatial-concatenation task in which a known signal is embedded within the denoising target, such as reference-based inpainting or outpainting, may benefit from the same decoupling principles.

Acknowledgement

This research was supported by the Regional Innovation System & Education (RISE) program through the Gyeongbuk RISE Center, funded by the Ministry of Education (MOE) and the Gyeongsangbuk-do, Republic of Korea. (2026-RISE-15-119)

Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. In: Int. Conf. Learn. Represent. (2018)
2. Bookstein, F.: Principal warps: thin-plate splines and the decomposition of deformations. IEEE Trans. Pattern Anal. Mach. Intell. **11**(6), 567–585 (1989)
3. Choi, S., Park, S., Lee, M., Choo, J.: Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 14131–14140 (2021)
4. Choi, Y., Kwak, S., Lee, K., Choi, H., Shin, J.: Improving diffusion models for authentic virtual try-on in the wild. In: Eur. Conf. Comput. Vis. pp. 206–235. Springer (2024)
5. Chong, Z., Dong, X., Li, H., Zhang, S., Zhang, W., Zhao, H., zhang, X., Jiang, D., Liang, X.: CatVTON: Concatenation is all you need for virtual try-on with diffusion models. In: Int. Conf. Learn. Represent. (2025)
6. Dube, D.R.: Enhanced virtual try on for clothing using deep learning. International Journal for Research in Applied Science and Engineering Technology (2024), `https://www.ijraset.com/best-journal/enhanced-virtual-try-on-for-clothing-using-deep-learning`, accessed: 2026-03-05
7. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Rombach, R.: Scaling rectified flow transformers for high-resolution image synthesis. In: Int. Conf. Mach. Learn. pp. 12606–12633 (2024)
8. Fele, B., Lampe, A., Peer, P., Struc, V.: C-vton: Context-driven image-based virtual try-on network. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 3144–3153 (2022)
9. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: Int. Conf. Learn. Represent. (2023)
10. Gong, K., Liang, X., Zhang, D., Shen, X., Lin, L.: Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 932–940 (2017)
11. 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)
12. 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: ACM Int. Conf. Multimedia. pp. 7599–7607 (2023)
13. Güler, R.A., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7297–7306 (2018)
14. Han, X., Hu, X., Huang, W., Scott, M.R.: Clothflow: A flow-based model for clothed person generation. In: Int. Conf. Comput. Vis. pp. 10471–10480 (2019)
15. Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S.: Viton: An image-based virtual try-on network. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7543–7552 (2018)
16. 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. In: Adv. Neural Inform. Process. Syst. pp. 6626–6637 (2017)
17. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural Inform. Process. Syst. **33**, 6840–6851 (2020)

18. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021)
19. Islam, T., Miron, A., Liu, X., Li, Y.: Deep learning in virtual try-on: A comprehensive survey. IEEE Access **12**, 29475–29502 (2024)
20. Ju, X., Liu, X., Wang, X., Bian, Y., Shan, Y., Xu, Q.: Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In: Eur. Conf. Comput. Vis. pp. 150–168. Springer (2024)
21. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. Adv. Neural Inform. Process. Syst. **35**, 26565–26577 (2022)
22. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4401–4410 (2019)
23. Kim, J., Gu, G., Park, M., Park, S., Choo, J.: Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 8176–8185 (2024)
24. 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. In: Int. Conf. Comput. Vis. pp. 16026–16036 (October 2025)
25. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: Int. Conf. Learn. Represent. (2014)
26. Lee, S., Gu, G., Park, S., Choi, S., Choo, J.: High-resolution virtual try-on with misalignment and occlusion-handled conditions. In: Eur. Conf. Comput. Vis. pp. 204–219. Springer (2022)
27. Lee, S., Kwak, J.g.: Voost: A unified and scalable diffusion transformer for bidirectional virtual try-on and try-off. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–11 (2025)
28. Li, P., Xu, Y., Wei, Y., Yang, Y.: Self-correction for human parsing. IEEE Trans. Pattern Anal. Mach. Intell. **44**(6), 3260–3271 (2020)
29. Li, Y., Zhao, H., Zhou, J., Xu, G., Hu, T., Chen, G., Wang, H.: A timestep-adaptive frequency-enhancement framework for diffusion-based image super-resolution. In: IJCAI. pp. 1503–1511 (2025)
30. Liang, X., Liu, S., Shen, X., Yang, J., Liu, L., Dong, J., Lin, L., Yan, S.: Deep human parsing with active template regression. IEEE Trans. Pattern Anal. Mach. Intell. **37**(12), 2402–2414 (2015)
31. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: Int. Conf. Learn. Represent. (2018)
32. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 11461–11471 (2022)
33. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDEdit: Guided image synthesis and editing with stochastic differential equations. In: Int. Conf. Learn. Represent. (2022)
34. Morelli, D., Baldrati, A., Cartella, G., Cornia, M., Bertini, M., Cucchiara, R.: Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. In: ACM Int. Conf. Multimedia. pp. 8580–8589 (2023)
35. Morelli, D., Fincato, M., Cornia, M., Landi, F., Cesari, F., Cucchiara, R.: Dress code: High-resolution multi-category virtual try-on. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2231–2235 (2022)

36. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Trans. Mach. Learn Res. (2024)
37. Park, Y.H., Kwon, M., Choi, J., Jo, J., Uh, Y.: Understanding the latent space of diffusion models through the lens of riemannian geometry. Adv. Neural Inform. Process. Syst. **36**, 24129–24142 (2023)
38. Parmar, G., Zhang, R., Zhu, J.Y.: On aliased resizing and surprising subtleties in gan evaluation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 11410–11420 (2022)
39. 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. In: Int. Conf. Learn. Represent. (2024)
40. 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: Int. Conf. Mach. Learn. pp. 8748–8763. PMLR (2021)
41. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10684–10695 (2022)
42. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
43. Sekri, K., Bouzaabia, O., Rzem, H., Juárez-Varón, D.: Effects of virtual try-on technology as an innovative e-commerce tool on consumers' online purchase intentions. European Journal of Innovation Management **28**(8), 4041–4060 (2025)
44. Stracke, N., Baumann, S.A., Bauer, K., Fundel, F., Ommer, B.: Cleandift: Diffusion features without noise. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 117–127 (2025)
45. Wang, B., Zheng, H., Liang, X., Chen, Y., Lin, L., Yang, M.: Toward characteristic-preserving image-based virtual try-on network. In: Eur. Conf. Comput. Vis. pp. 589–604 (2018)
46. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
47. Xie, Z., Huang, Z., Dong, X., Zhao, F., Dong, H., Zhang, X., Zhu, F., Liang, X.: Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 23550–23559 (2023)
48. Xie, Z., Huang, Z., Zhao, F., Dong, H., Kampffmeyer, M., Liang, X.: Towards scalable unpaired virtual try-on via patch-routed spatially-adaptive gan. Adv. Neural Inform. Process. Syst. **34**, 2598–2610 (2021)
49. Xu, Y., Gu, T., Chen, W., Chen, C.: Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. In: AAAI (2025)
50. Yang, H., Zhang, R., Guo, X., Liu, W., Zuo, W., Luo, P.: Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7850–7859 (2020)

51. Yang, Z., Jiang, Z., Li, X., Zhou, H., Dong, J., Zhang, H., Du, Y.: D4-vton: Dynamic semantics disentangling for differential diffusion based virtual try-on. In: Eur. Conf. Comput. Vis. pp. 36–52. Springer (2024)
52. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
53. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Int. Conf. Comput. Vis. pp. 3836–3847 (2023)
54. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 586–595 (2018)
55. Zhou, J., Ding, T., Chen, T., Jiang, J., Zharkov, I., Zhu, Z., Liang, L.: Dream: Diffusion rectification and estimation-adaptive models. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 8342–8351 (2024)
56. Zhou, Z., Wang, S., Ge, T., Jiang, Y.: A high-resolution image-based virtual try-on system in taobao e-commerce scenario. In: ACM Int. Conf. Multimedia. pp. 6970–6972 (2022)
57. 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: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2491–2501 (2025)
58. Zhu, L., Yang, D., Zhu, T., Reda, F., Chan, W., Saharia, C., Norouzi, M., Kemelmacher-Shlizerman, I.: Tryondiffusion: A tale of two unets. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4606–4615 (2023)