INVERSE VIRTUAL TRY-ON: GENERATING MULTI-CATEGORY PRODUCT-STYLE IMAGES FROM CLOTHED INDIVIDUALS

Anonymous authors
Paper under double-blind review

Figure 1: Visual results produced by our proposed text-enhanced multi-category virtual try-off architecture, *i.e.*, TEMU-VTOFF. Given a clothed input person image, the proposed model reconstructs the clean, in-shop version of the worn garment. Our model handles various garment types and preserves both structural fidelity and fine-grained textures, even under occlusions and complex poses, thanks to its multimodal attention and garment-alignment design.

ABSTRACT

Virtual try-on (VTON) has been widely explored for rendering garments onto person images, while its inverse task, virtual try-off (VTOFF), remains largely overlooked. VTOFF aims to recover standardized product images of garments directly from photos of clothed individuals. This capability is of great practical importance for e-commerce platforms, large-scale dataset curation, and the training of foundation models. Unlike VTON, which must handle diverse poses and styles, VTOFF naturally benefits from a consistent output format in the form of flat garment images. However, existing methods face two major limitations: *(i)* exclusive reliance on visual cues from a single photo often leads to ambiguity, and *(ii)* generated images usually suffer from loss of fine details, limiting their real-world applicability. To address these challenges, we introduce **TEMU-VTOFF**, a **T**ext-**E**nhanced **MU**lti-category framework for **VTOFF**. Our architecture is built on a dual DiT-based backbone equipped with a multimodal attention mechanism that jointly exploits image, text, and mask information to resolve visual ambiguities and enable robust feature learning across garment categories. To explicitly mitigate detail degradation, we further design an alignment module that refines garment structures and textures, ensuring high-quality outputs. Extensive experiments on VITON-HD and Dress Code show that TEMU-VTOFF achieves new state-of-the-art performance, substantially improving both visual realism and consistency with target garments. Code and models will be released to foster future research.

1 INTRODUCTION

Unlike virtual try-on (VTON), whose goal is to dress a given clothing image on a target person image, in this paper, we focus exactly on the opposite, virtual try-off (VTOFF), whose purpose is to generate standardized product images from real-world clothed individual photos. Compared to

VTON, which often struggles with the ambiguity and diversity of valid outputs, such as stylistic variations in how a garment is worn, VTOFF benefits from a clearer output objective: *reconstructing a consistent, lay-down-style image of the garment*. This reversed formulation facilitates a more objective evaluation of garment reconstruction quality.

The fashion industry, a trillion-dollar global market, is increasingly integrating AI and computer vision to optimize product workflows and enhance user experience. VTOFF, in this context, offers substantial value: it enables the automatic generation of tiled product views, which are essential for tasks such as image retrieval, outfit recommendation, and virtual shopping. However, acquiring such lay-down images is expensive and time-consuming for retailers. VTOFF provides a scalable alternative by leveraging images of garments worn by models or customers, transforming them into standardized catalog views through image-to-image translation techniques.

Despite the success of GANs (Goodfellow et al., 2014) and Latent Diffusion Models (LDMs) (Rombach et al., 2022) in image translation tasks (Siarohin et al., 2019; Ren et al., 2023b; Isola et al., 2017; Tumanyan et al., 2023), current VTOFF solutions face notable limitations. Existing models (Velioglu et al., 2024; Xarchakos & Koukopoulos, 2024) struggle to accurately reconstruct catalog images from dressed human inputs. This limitation arises from a fundamental architectural mismatch: these approaches repurpose VTON pipelines by merely reversing the input-output roles, without addressing the unique challenges of the VTOFF task. Moreover, the high visual variability of real-world images – due to garment wear category (*e.g.*, upper-body), pose changes, and occlusions – makes it difficult for these models to robustly extract garment features while preserving fine-grained patterns. On the opposite side, we design a dedicated architecture tailored for the VTOFF task.

Recent advances in diffusion models demonstrate that DiT-based architectures (Peebles & Xie, 2023), especially when combined with flow-matching objectives (Lipman et al., 2023), surpass traditional U-Net and DDPM-based approaches (Rombach et al., 2022). Inspired by these findings, we propose **TEMU-VTOFF**, a **T**ext-**E**nhanced **MU**lti-category **V**irtual **T**ry-**OFF** architecture based on a dual-DiT framework. Specifically, we exploit the representational strength of DiT in two distinct ways: *(i)* the first Transformer component focuses on extracting fine-grained garment features from complex, detail-rich person images; and *(ii)* the second DiT is specialized for generating the clean, in-shop version of the garment. To support this design, we further adapt the base DiT architecture to accommodate the task-specific input modalities. To further enhance alignment, we introduce an external garment aligner module and a novel supervision loss that leverages clean garment references as guidance, further improving quality of generated images.

Our contribution can be summarized as follows:

- **Multi-Category Try-Off.** We present a unified framework capable of handling multiple garment types (upper-body, lower-body, and full-body clothes) without requiring category-specific pipelines.
- **Multimodal Hybrid Attention.** We introduce a novel attention mechanism that integrates garment textual descriptions into the generative process by linking them with person-specific features. This helps the dual-DiT architecture synthesize the garments more accurately.
- **Garment Aligner Module.** We design a lightweight aligner that conditions generation on clean garment images, replacing conventional denoising objectives. This leads to better alignment consistency on the overall dataset and preserves more precise visual retention.
- Extensive experiments on the Dress Code and VITON-HD datasets demonstrate that TEMU-VTOFF outperforms prior methods in both the quality of generated images and alignment with the target garment, highlighting its strong generalization capabilities.

2 Related Work

Virtual Try-On. As one of the most popular tasks within the fashion domain, VTON has been widely studied over the past decades by the computer vision and graphics communities due to its interesting challenges and the practical potential (Bai et al., 2022; Cui et al., 2021; Fele et al., 2022; Ren et al., 2023a). Existing methods are broadly categorized into warping-based (Chen et al., 2023; Xie et al., 2023; Yan et al., 2023) and warping-free approaches (Zhu et al., 2023; Morelli et al., 2023; Baldrati et al., 2023; Zeng et al., 2024; Chong et al., 2025), with a growing shift from GAN-based (Goodfellow et al., 2020) to diffusion-based (Ho et al., 2020; Song et al., 2021) frameworks.

VITON (Han et al., 2018), CP-VTON (Wang et al., 2018), and their variants improve garment alignment and synthesis quality, but often produce artifacts due to imperfect warping. To mitigate this, warping-free methods leverage diffusion models to bypass explicit deformation (Zhu et al., 2023; Morelli et al., 2023; Xu et al., 2025; Choi et al., 2024) employing modified cross-attention or self-attention to directly condition generation on garment features. However, these pre-trained encoders tend to lose fine-grained texture details, prompting methods like StableVITON (Kim et al., 2024) to introduce dedicated garment encoders and attention mechanisms, albeit at a higher computational cost. Lately, DiT-based works (Jiang et al., 2025; Zhu et al., 2024) show the benefits of Transformer-based diffusion models for high-fidelity garment to person transfer. Finally, some models adopt more elaborate conditioning strategies. For instance, LOTS introduces a pair-former module for handling multiple inputs (Girella et al., 2025), while LEFFA learns a flow field from averaged cross-attention maps and employs learnable tokens to stabilize attention values (Zhou et al., 2025). While most works focus on generating dressed images from separate person and garment inputs, the inverse problem (*i.e.*, reconstructing clean garment representations from worn images) remains underexplored.

Virtual Try-Off. While VTON has been extensively studied for synthesizing images of a person wearing a target garment, the recently proposed VTOFF task shifts the focus toward garment-centric reconstruction, aiming to extract a clean, standardized image of a garment worn by a person. TryOffDiff (Velioglu et al., 2024) introduces this task by leveraging a diffusion-based model with SigLIP (Zhai et al., 2023) conditioning to recover high-fidelity garment images. Building on this direction, TryOffAnyone (Xarchakos & Koukopoulos, 2024) addresses the generation of tiled garment images from dressed photos for applications like outfit composition and retrieval. By integrating garment-specific masks and simplifying the Stable Diffusion pipeline through selective Transformer tuning, it achieves both quality and efficiency. In both cases, these works have been designed for single-category scenarios, thus limiting their potential application to generate wider, more diverse data collections. Recent efforts have begun to address these limitations. MGT (Velioglu et al., 2025) extends VTOFF to multi-category scenarios by incorporating class-specific embeddings to handle diverse clothing types within a unified model. More ambitious approaches aim to unify both VTON and VTOFF within a single framework. Voost (Lee & Kwak, 2025) proposes a single diffusion transformer to learn both tasks, while One Model For All (Liu et al., 2025) introduces a partial diffusion mechanism to achieve a similar goal. On a different line, Any2AnyTryon (Guo et al., 2025) is not a native VTOFF method, but it leverages a LoRA-based module (Hu et al., 2022) to fine-tune FLUX (Labs, 2024) for this task. Though these works collectively reflect a growing shift from person-centric synthesis to garment-centric understanding, there are still limitations like frequent garment structural artifacts (*e.g.*, in shape, neckline, waist) and on colors and textures of generated outputs. We hypothesize that this mismatch is due to a too generic architectural choice, not tailored for the specific needs of the VTOFF setting. In this work, we focus on existing VTOFF open problems, such as multi-category adaptation, occlusions, and complex human poses, and propose a novel VTOFF-specific architecture enhanced with text and fine-grained mask conditioning and optimized with a garment aligner component that can improve the quality of generated garments.

Conditioning Methods in Diffusion Models. To overcome the limitations of text-only conditioning, many schemes leverage additional visual inputs such as segmentation maps, bounding boxes, poses, and points (Sun et al., 2024; Li et al., 2023; Chen et al., 2024; Nie et al., 2024; Lin et al., 2024; Wang et al., 2024). Prominent methods like ControlNet (Zhang et al., 2023) and T2I-Adapter (Mou et al., 2024) inject spatial conditions via auxiliary networks, while IP-Adapter (Ye et al., 2023) uses separate attention branches more suited to U-Nets than DiTs. Other works focus on unifying multiple conditions, either through modular controllers like Uni-ControlNet (Zhao et al., 2023) and dedicated adapters (Lin et al., 2025), or by concatenating visual embeddings directly into the Transformer input sequence (Tan et al., 2025; Wang et al., 2025; Xiao et al., 2025). Although these methods are effective for general personalization tasks (*i.e.*, placing an object from one image into another), they lack the fine-grained conditioning mechanism necessary to extract specific garment data from person images, a gap we address to unlock the VTOFF task.

3 METHODOLOGY

Preliminaries. The latest diffusion models are a family of generative architectures that corrupt a ground-truth image $\boldsymbol{z}_0$ according to a flow-matching schedule (Lipman et al., 2023) defined as

$$\boldsymbol{z}_t = (1-t)\boldsymbol{z}_0 + t\epsilon_t \quad \epsilon \sim \mathcal{N}(0,1), \quad t \in [0,1]. \tag{1}$$

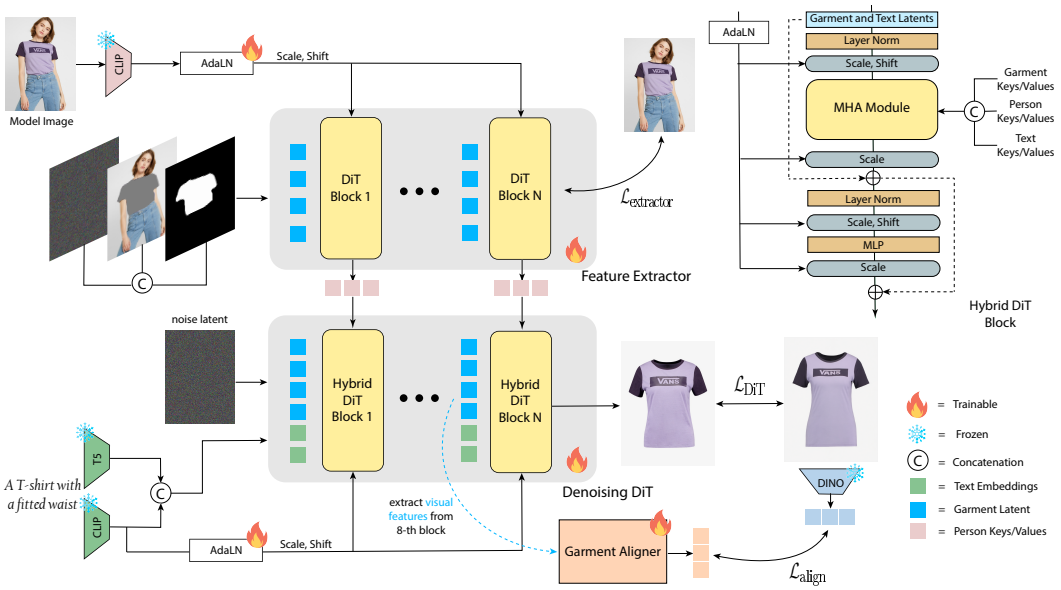


Figure 2: Overview of our method. The feature extractor F_E processes spatial inputs (noise, masked image, binary mask), and global inputs (model image via AdaLN). The intermediate keys and values $\boldsymbol{K}^l_{\text{extractor}}, \boldsymbol{V}^l_{\text{extractor}}$ are injected into the corresponding hybrid blocks of the garment generator F_D. Then, the main DiT model generates the final garment leveraging the proposed MHA module. We align our model with a diffusion loss for the noise estimate and an alignment loss with clean, DINOv2 features of the target garment.

Then, a diffusion model estimates back the injected noise ϵ_t through a Diffusion Transformer (DiT) (Peebles & Xie, 2023), obtaining a prediction $\hat{z}_0$. In Stable Diffusion 3 (SD3) (Esser et al., 2024), the 16-channel latent $\boldsymbol{z}_t \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 16}$ is obtained projecting the original RGB image $\boldsymbol{x} \in \mathbb{R}^{H \times W \times 3}$ with a variational autoencoder $\mathcal{E}$ (Kingma & Welling, 2013), obtaining $\boldsymbol{z} = \mathcal{E}(\boldsymbol{x})$, with H, W being height and width of the image, and $f = 8$ the spatial compression ratio of the autoencoder. Finally, the model is trained according to an MSE loss function $\mathcal{L}_{\text{diff}}$:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\boldsymbol{z}_0, \epsilon_t, t}\left[\|\epsilon_t - \epsilon_\theta(\boldsymbol{z}_t, t)\|^2\right]. \tag{2}$$

Overview. An overview of our method is shown in Fig. 2. The objective is to generate an in-shop version of the garment worn by the person. A critical design choice lies in processing the dressed person image so as to extract meaningful information for injection into the denoising process. To this end, we adopt a dual-DiT architecture, built upon SD3, with the two models assigned to complementary roles. Firstly, we design the first DiT as a feature extractor F_E that encodes the model image $\boldsymbol{x}_{\text{model}}$ and outputs its intermediate layer features at timestep $t = 0$ and not from subsequent timesteps, as we are interested in extracting clean features from F_E. This block is trained with a diffusion loss to generate the person image. Once trained, this model outputs meaningful key and value features of the dressed person. Secondly, the main DiT generates the garment $\boldsymbol{x}_g$ leveraging the intermediate features from F_E in a modified textual-enhanced attention module.

3.1 DiT Feature Extractor

At inference time, the only available input is the clothed person image $\boldsymbol{x}_{\text{model}} \in \mathbb{R}^{H \times W \times 3}$, from which we also extract the mask. To encode this information, we compute the visual projection

$$\boldsymbol{e}^v_{\text{pool}} = \text{CLIP}(\boldsymbol{x}_{\text{model}}) \in \mathbb{R}^{2048},$$

which is then used to modulate the latent $\boldsymbol{z}_t$ through the AdaLN-estimated scale γ and shift β as follows:

$$\begin{aligned} \boldsymbol{y}_t &= \text{MLP}(t, \boldsymbol{e}^v_{\text{pool}}), \\ \boldsymbol{z}_t &\leftarrow \gamma(\boldsymbol{y}_t)\boldsymbol{z}_t + \beta(\boldsymbol{y}_t). \end{aligned} \tag{3}$$

Existing VTON approaches rely on two visual inputs: the target garment and the person. In our case, however, the model can rely only on person features, from which it is more complex to extract

the garment features. This shortcoming makes the CLIP vector $\boldsymbol{e}^{v}_{\text{pool}}$ the real bottleneck in a unified architecture setting, as the CLIP projection alone is too coarse to properly encode this information.

To address this, we propose introducing a dedicated feature extractor F_E, allowing F_D to concentrate exclusively on the garment generation task. The architecture of F_E mirrors that of the main SD3 DiT module F_D, with the only difference being its input layer, which is adapted to handle additional visual inputs in the channel dimension. The inputs are a global input with the person image $\boldsymbol{x}_{model} \in \mathbb{R}^{H\times W\times 3}$ encoded as $\boldsymbol{e}^{v}_{\text{pool}}$ and leveraged by the modulation layers of F_E, and a local spatial input as the channel-wise concatenation $\boldsymbol{z}'_t = [\boldsymbol{z}_t, M, \boldsymbol{x}_M] \in \mathbb{R}^{h\times w\times 33}$ of the latent $\boldsymbol{z}_t$, the encoded latent of the masked person image $\boldsymbol{x}_M = \mathcal{E}(I_M) \in \mathbb{R}^{h\times w\times 16}$, and the interpolated binary mask $M \in \mathbb{R}^{h\times w\times 1}$ encoded through the Transformer projector $\mathcal{P} : \mathbb{R}^{h\times w\times 33} \to \mathbb{R}^{S\times d}$, with S as sequence length and d as embedding dimension.

This design choice is central to our method, as each layer output of the feature extractor F_E retains meaningful intermediate representations of both the person and the garment. Leveraging these features offers three key advantages: *(i)* instead of the collapsed CLIP representation, we obtain expanded features of dimension $S \times d$; *(ii)* the L layers of F_E capture information at multiple granularities, progressing from coarse to fine (Avrahami et al., 2025; Skorokhodov et al., 2025), so that each layer l conveys a different level of detail about the same image; *(iii)* since F_E shares the same architecture as F_D, the features extracted at layer l from F_E are naturally better aligned with those of F_D. Motivated by these considerations, we extract the keys $\boldsymbol{K}^{l}_{\text{extractor}}$ and values $\boldsymbol{V}^{l}_{\text{extractor}}$ from every layer l of F_E.

3.2 Dual-DiT Text-Enhanced Garment Try-Off

Without loss of generality, we will omit the index l when referring to $\boldsymbol{Q}^l$ $\boldsymbol{K}^l$ and $\boldsymbol{V}^l$ of F_E and F_D, since the conditioning scheme is applied uniformly across all layers. Given the extracted features $\boldsymbol{K}_{\text{extractor}}$ and $\boldsymbol{V}_{\text{extractor}}$, we propose to modify the SD3 attention scheme to incorporate such information, leading to our Multimodal Hybrid Attention (MHA).

Multimodal Hybrid Attention. Our new module seamlessly mix text information, latent features of the denoising DiT, and intermediate features from F_E. Inspired by the key findings in SD3 (Esser et al., 2024), we concatenate the text features with the visual inputs along the sequence length dimension, thus obtaining:

$$\boldsymbol{Q} = [\boldsymbol{Q}_{\boldsymbol{z}_t}, \boldsymbol{Q}_{\text{text}}] \quad \boldsymbol{K} = [\boldsymbol{K}_{\boldsymbol{z}_t}, \boldsymbol{K}_{\text{extractor}}, \boldsymbol{K}_{\text{text}}] \quad \boldsymbol{V} = [\boldsymbol{V}_{\boldsymbol{z}_t}, \boldsymbol{V}_{\text{extractor}}, \boldsymbol{V}_{\text{text}}]. \tag{4}$$

This module allows the features $\boldsymbol{Q}_{\text{text}}$ to attend both the latent projection $\boldsymbol{K}_{\boldsymbol{z}_t}$ and the extractor features $\boldsymbol{K}_{\text{extractor}}$. The resulting attention matrix $\boldsymbol{A}_{\text{MHA}}$ captures three key interactions: (i) $\boldsymbol{A}_{\text{text}\leftrightarrow\boldsymbol{z}_t}$, preserving pre-trained alignment between language and latent image tokens, (ii) $\boldsymbol{A}_{\boldsymbol{z}_t\leftrightarrow\text{extractor}}$, facilitating transfer between the input garment and the person representation, and (iii) $\boldsymbol{A}_{\text{text}\leftrightarrow\text{extractor}}$, grounding the text in the structural features provided by the extractor.

Text embeddings are constructed via the concatenation of CLIP (Radford et al., 2021)[1] and T5 (Raffel et al., 2020) encoders applied to the input caption c as follows:

$$\boldsymbol{e}_{\text{text}} = [\text{CLIP}(c), \text{T5}(c)], \quad \text{with } \boldsymbol{e}_{\text{text}} \in \mathbb{R}^{77\times 4096}. \tag{5}$$

Now we pose a relevant question: is it possible to disambiguate the garment category from the mask alone? A mask input can improve multi-category handling by acting as a *hard* discriminator between two garments, in contrast to text, which acts as a *soft* discriminator since it does not directly indicate the pixels occupied by the target garment. Therefore, the mask can help to visually force the model to retain only upper- or lower-body information but it can not tell much about the appearance of a garment, because it is highly warped together with the person, resulting in visual artifacts. Textual information is critical, together with mask information, to extract the category information of the garment. To address this, we decide to use also the global conditioning scheme provided by AdaLN (Huang & Belongie, 2017) in SD3. As shown in previous works (Garibi et al., 2025), these layers can be successfully leveraged to adapt "appearance" or "style" information into existing Transformer-based architectures. For this reason, we extract a pooled textual representation $\boldsymbol{e}_{\text{pool}} \in \mathbb{R}^{2048}$ of CLIP textual features of the caption c and inject them into the model through

[1]Following SD3, we consider the combined embedding from CLIP ViT-L and Open-CLIP bigG/14.

the modulation layers, following Eq. 3. The pooled vector $\boldsymbol{e}_{\text{pool}} \in \mathbb{R}^{2048}$ encapsulates a coarser representation than the full textual embeddings $\boldsymbol{e}_{\text{text}} \in \mathbb{R}^{77\times4096}$, thus being suitable for high-level information conditioning.

Training. We employ a two-stage training procedure: we train the module F_E alone, detached from the dual DiT F_D, according to the diffusion loss L_{diff} defined as follows:

$$\mathcal{L}_{\text{extractor}} = \mathbb{E}_{\boldsymbol{z}_0,\epsilon_t,t}\left[\|\epsilon_t - F_E(\boldsymbol{z}'_t, \boldsymbol{x}_{\text{model}}, t)\|^2\right]. \tag{6}$$

Then, we train the main DiT module (*i.e.*, F_D) following a diffusion loss with multiple conditioning signals:

$$\mathcal{L}_{\text{DiT}} = \mathbb{E}_{\boldsymbol{z}_g,\epsilon_t,t}\left[\|\boldsymbol{z}_g - F_D(\boldsymbol{z}_t, \boldsymbol{e}_{\text{pool}}, F_E(\boldsymbol{z}'_0, \boldsymbol{x}_{\text{model}}, 0), t)\|^2\right], \tag{7}$$

with $F_E(\boldsymbol{z}'_0, \boldsymbol{x}_{\text{model}}, 0)$ being the list of keys and values extracted from F_E at timestep $t = 0$. We extract this list from F_E at $t = 0$ and re-use them in F_D for all subsequent timesteps, as we want to use key/values from clean data.

3.3 Garment Aligner

While our model is effective at generating realistic and structurally coherent garments, we observe occasional failures in preserving high-frequency details such as fine-grained textures and logos. We hypothesize two primary contributing factors: (i) the diffusion loss $\mathcal{L}_{\text{diff}}$, defined in the noise space, optimizes over perturbed latents rather than directly over image-space reconstructions, limiting its sensitivity to fine-grained patterns; and (ii) the inherent generation dynamics of diffusion models, where errors introduced in early timesteps – typically encoding low-frequency content – can accumulate and degrade the fidelity of high-frequency details in later stages. To mitigate this, we draw inspiration from REPA (Yu et al., 2025), and propose to explicitly align the internal feature representation of our DiT with that of a pre-trained vision encoder. Specifically, we encourage patch-wise consistency between the eighth Transformer block of our main DiT model F_D and the corresponding features extracted from DINOv2 (Oquab et al., 2023).

Formally, let $\boldsymbol{h}_{\text{DiT}} \in \mathbb{R}^{3072\times d}$ denote the token sequence obtained from the eighth Transformer block of the DiT decoder F_D, corresponding to a 64×48 patch grid with embedding dimension d. Separately, let $\boldsymbol{h}_{\text{enc}} \in \mathbb{R}^{1024\times d'}$ be the 32×32 token grid extracted from a frozen DINOv2 encoder, with embedding dimension d' (where $d' \neq d$). To bridge this mismatch, we introduce a lightweight garment aligner module composed of a convolutional neural network $\phi_{\text{CNN}} : \mathbb{R}^{64\times48\times d} \to \mathbb{R}^{32\times32\times d'}$ which is used to downsample the spatial token grid while preserving local structure and to project the token embeddings into the DINOv2 feature space. The aligned tokens are defined as $\tilde{\boldsymbol{h}}_{\text{DiT}} = \phi_{\text{CNN}}(\boldsymbol{h}_{\text{DiT}}) \in \mathbb{R}^{1024\times d'}$.

We then enforce feature-level consistency via a cosine similarity loss:

$$\mathcal{L}_{\text{align}} = -\mathbb{E}_{\boldsymbol{z}_g,\epsilon_t,t}\left[\frac{1}{N}\sum_{i=1}^{N}\cos\left(\tilde{\boldsymbol{h}}_i^{\text{DiT}}, \boldsymbol{h}_i^{\text{enc}}\right)\right], \tag{8}$$

where $\tilde{\boldsymbol{h}}_i^{\text{DiT}}$ and $\boldsymbol{h}_i^{\text{enc}}$ are the i-th aligned and reference tokens, respectively, i is the patch index, N is the total number of tokens, and cos is the cosine similarity.

Overall Loss Function. The garment aligner is applied in the second stage of our training. Our final training objective combines the standard diffusion loss $\mathcal{L}_{\text{DiT}}$ with the garment alignment loss $\mathcal{L}_{\text{align}}$ previously introduced. The overall objective is thus defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{DiT}} + \lambda \cdot \mathcal{L}_{\text{align}}, \tag{9}$$

where λ is a hyperparameter that balances the contribution of the two loss components.

4 Experiments

4.1 Comparison with the State of the Art

We conduct our experiments using two publicly available fashion datasets: VITON-HD (Choi et al., 2021) and Dress Code (Morelli et al., 2022). VITON-HD contains only upper-body garments and

Table 1: Quantitative results on the Dress Code dataset, considering both the entire test set and the three category-specific subsets. ↑ indicates higher is better, ↓ lower is better.

	All					Upper-Body				
Method	**SSIM ↑**	**LPIPS ↓**	**DISTS ↓**	**FID ↓**	**KID ↓**	**SSIM ↑**	**LPIPS ↓**	**DISTS ↓**	**FID ↓**	**KID ↓**
TryOffDiff	-	-	-	-	-	76.59	40.62	29.04	37.97	17.30
Any2AnyTryon	77.56	35.17	25.17	12.32	3.65	76.61	38.99	25.78	15.77	3.22
MGT	**77.77**	35.37	27.28	13.47	5.28	**76.77**	39.70	28.13	19.49	6.87
TEMU-VTOFF *(Ours)*	75.95	**31.46**	**18.66**	**5.74**	**0.65**	74.54	**35.48**	**19.75**	**10.94**	**0.76**
	Lower-Body					**Dresses**				
Method	**SSIM ↑**	**LPIPS ↓**	**DISTS ↓**	**FID ↓**	**KID ↓**	**SSIM ↑**	**LPIPS ↓**	**DISTS ↓**	**FID ↓**	**KID ↓**
Any2AnyTryon	**78.15**	34.72	25.87	30.06	12.01	77.93	31.80	23.86	19.20	6.27
MGT	77.29	36.31	28.00	25.98	9.64	79.26	30.11	25.70	19.09	5.74
TEMU-VTOFF *(Ours)*	73.94	**34.60**	**19.57**	**13.83**	**2.04**	**79.39**	**24.32**	**16.67**	**11.29**	**0.59**

Figure 3: Qualitative comparison on the Dress Code dataset between images generated by TEMU-VTOFF and those generated by competitors.

represents a single-category setting, while Dress Code includes multiple categories (*i.e.*, dresses, upper-body, and lower-body garments) enabling evaluation of the generalization capabilities of our methods across diverse garment types. To evaluate the proposed TEMU-VTOFF architecture, we use a combination of perceptual, structural, and distributional similarity metrics. Specifically, we report LPIPS (Zhang et al., 2018), SSIM (Wang et al., 2004), DISTS (Ding et al., 2020), FID (Parmar et al., 2022), and KID (Bińkowski et al., 2018). We compare our approach against recent VTOFF methods, including TryOffDiff (Velioglu et al., 2024), TryOffAnyone (Xarchakos & Koukopoulos, 2024), MGT (Velioglu et al., 2025), Voost (Lee & Kwak, 2025), One Model For All (Liu et al., 2025), and Any2AnyTryon (Guo et al., 2025). For TryOffAnyone, Voost, and One Model For All, we report the results only on the VITON-HD dataset because they have not been trained on the Dress Code dataset. Additionally, we retrain TryOffDiff on Dress Code using the official code and hyperparameters provided by the authors. Since TryOffDiff is not designed to handle multi-category garments, we report results only for the upper-body category.

Results on the Dress Code dataset. Table 1 reports the experimental results on the Dress Code dataset. As observed, our method outperforms existing state-of-the-art approaches across most evaluation metrics and garment categories. These results indicate that our approach is category-agnostic and benefits from the joint use of textual garment descriptions and fine-grained masks. Consequently, our model achieves a better perceptual quality and closer alignment with the ground-truth distribution compared to competing methods.

In Fig. 3, we provide qualitative results comparing TEMU-VTOFF with competitors. These examples highlight the challenges posed by the diverse set of categories in Dress Code. As shown, MGT and Any2AnyTryon frequently struggle to preserve key visual attributes such as color, texture, and shape. In contrast, our method is able to closely match the target garment across all categories, demonstrating a clear improvement in generation quality.

Results on the VITON-HD dataset. In Table 2, we report the quantitative results on VITON-HD. In this setting, TEMU-VTOFF sets a new state-of-the-art across the majority of metrics, achieving the best scores for DISTS, FID, and KID. This indicates a superior ability to reconstruct structural details and to match the distribution of the ground-truth images. Notably, One Model for All achieves a

Table 2: Quantitative results on the VITON-HD dataset. ↑ indicates higher is better, ↓ lower is better. † denotes results taken directly from the original papers.

Method	SSIM ↑	LPIPS ↓	DISTS ↓	FID ↓	KID ↓
TryOffDiff	75.53	39.56	25.53	17.49	5.30
TryOffAnyone	75.90	35.26	23.47	12.74	2.85
Any2AnyTryon	75.72	37.95	24.32	12.88	3.01
MGT†	**78.10**	36.30	24.70	21.90	8.90
Voost†	-	-	-	10.06	2.48
One Model for All†	-	**22.50**	19.20	9.12	1.49
TEMU-VTOFF *(Ours)*	77.21	28.44	**18.04**	**8.71**	**1.11**

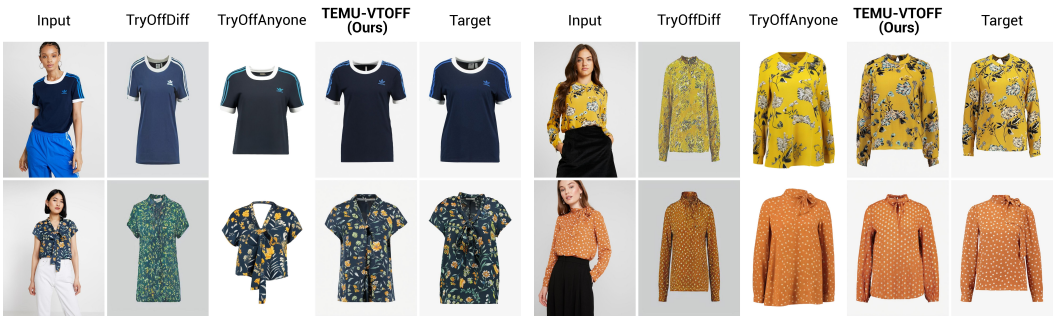


Figure 4: Qualitative comparison on the VITON-HD dataset between images generated by TEMU-VTOFF and those generated by competitors.

competitive LPIPS score, suggesting strong perceptual similarity. However, our method demonstrates a more robust and balanced performance, significantly outperforming all other approaches on the distributional FID and KID metrics, which are critical for assessing image realism and diversity.

Overall, our method achieves solid improvements on VITON-HD, although the performance gains are less pronounced than on Dress Code. This is expected, as VITON-HD focuses exclusively on upper-body garments and is therefore a simpler benchmark. In contrast, the diverse and multi-category nature of Dress Code, with dresses, skirts, and pants, highlights the advantages of our approach, where the joint use of textual descriptions and fine-grained masks proves critical for accurate garment reconstruction. Accordingly, the strengths of our method are most evident in complex, multi-category scenarios. A visual comparison on sample VITON-HD images is shown in Fig. 4, which further demonstrates the improved garment reconstruction quality of our proposed method.

4.2 Ablation Studies

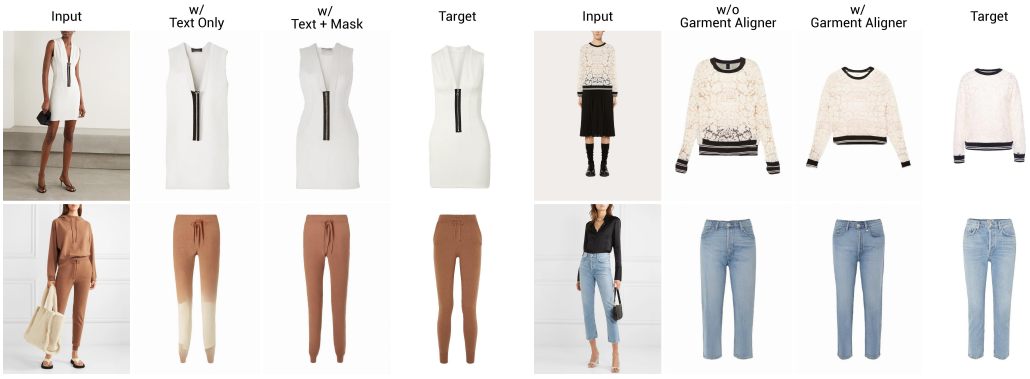
To assess the contribution of each component in our pipeline, we conduct a detailed ablation study on the Dress Code dataset reported in Table 3. Removing garment descriptions or fine-grained masks consistently reduces performance, with the largest drop when both are absent, confirming that masks act as spatial anchors while text provides complementary semantic and category-level cues. The best results are obtained when both inputs are present, highlighting their complementarity.

Further, we investigate the impact of our dual-stream DiT architecture by removing the feature extractor F_E. As shown, without the feature extractor, we experience a clear performance drop. In contrast, injecting $t = 0$ keys and values from F_E into the generator component through the proposed MHA operator enables richer, multi-scale conditioning, leading to better results. Finally, removing the garment aligner module reduces perceptual fidelity, particularly in categories with complex structures such as dresses, confirming that all designed components plays a critical role to the final performance.

To better understand the strength of each component proposed in our approach, we provide a visual comparison on the Dress Code dataset in Fig. 5. When our method relies exclusively on visual features from the person, without any textual guidance, it can struggle to resolve ambiguities in the garment design. This often leads to inaccuracies in key structural elements like the neckline, sleeve

Table 3: Ablation study of the proposed components on the Dress Code dataset.

	All					Upper-body		Lower-body		Dresses	
	SSIM ↑	LPIPS ↓	DISTS ↓	FID ↓	KID ↓	DISTS ↓	FID ↓	DISTS ↓	FID ↓	DISTS ↓	FID ↓
Effect of Text and Mask Conditioning											
w/o text and masks	71.04	39.68	25.20	9.63	3.17	23.71	19.75	65.85	49.19	20.12	15.47
w/o text modulation	73.88	34.63	22.54	7.75	1.52	24.02	13.48	24.33	18.13	19.27	13.30
w/o fine-grained masks	74.65	32.33	20.87	6.58	1.03	20.85	11.31	22.34	15.74	19.42	13.62
TEMU-VTOFF *(Ours)*	**75.95**	**31.46**	**18.66**	**5.74**	**0.65**	**19.75**	**10.94**	**19.57**	**13.83**	**16.67**	**11.29**
Effect of Dual-Stream DiT											
w/o feature extractor F_E	72.79	38.61	23.56	9.11	1.70	24.97	14.13	23.20	19.54	22.52	16.82
TEMU-VTOFF *(Ours)*	**76.01**	**30.84**	**20.63**	**5.91**	**0.78**	**21.77**	**11.26**	**22.26**	**14.22**	**17.86**	**11.86**
Effect of Garment Aligner Component											
w/o garment aligner	**76.01**	**30.84**	20.63	5.91	0.78	21.77	11.26	22.26	14.22	17.86	11.86
TEMU-VTOFF *(Ours)*	75.95	31.46	**18.66**	**5.74**	**0.65**	**19.75**	**10.94**	**19.57**	**13.83**	**16.67**	**11.29**



(a) Evaluation of mask and text joint impact.

(b) Evaluation of garment aligner impact.

Figure 5: Qualitative comparisons validating the effectiveness of the proposed components on the Dress Code dataset.

length, or overall fit. The introduction of a textual description addresses this by acting as crucial structural guidance. This enables the model to capture the fundamental parts of the garment, ensuring the generated item correctly reflects the intended type and style. Subsequently, the fine-grained mask provides a precise spatial boundary, enforcing a clean silhouette and sharp edges, which improves the overall shape and contour of the garment. Finally, the garment aligner further improves the visual fidelity by encouraging the reconstruction of high-frequency details. This results in improved textures and more accurate patterns, ensuring that the final generated garment is not only structurally correct but also rich in fine-grained detail.

Limitations. While our method shows strong performance and generalization, some limitations remain. First, it struggles to reconstruct fine-grained details such as logos or printed text, partly due to the SD3 backbone. Second, performance on lower-body garments is less reliable than for upper-body garments and dresses, likely due to class imbalance in the Dress Code dataset. Additional discussion and failure cases are included in the supplementary material.

5 Conclusion

We have presented TEMU-VTOFF, a novel architecture that pushes the boundaries of VTOFF for complex, multi-category scenarios. While existing methods often struggle with detail preservation and accurate reconstruction across diverse garment types, our approach is specifically designed to overcome these limitations. We achieve this through a novel dual-DiT framework that leverages multimodal hybrid attention to effectively fuse information from the person, the garment, and textual descriptions. To further enhance realism, our proposed garment aligner module refines fine-grained textures and structural details. The effectiveness of our method is validated by state-of-the-art performance on standard VTOFF benchmarks, demonstrating its robustness in generating high-fidelity, catalog-style images.

Ethics Statement

Our method addresses the VTOFF task by generating flat, in-shop garment images from photos of dressed individuals. This enables a novel form of data augmentation in the fashion domain, allowing clean garment representations to be synthesized without manual segmentation or dedicated photoshoots. By bridging the gap between worn and catalog-like appearances, our approach can improve scalability for fashion datasets and support downstream applications such as retrieval, recommendation, and virtual try-on. However, as with any generative technology, there are important ethical and legal considerations. In particular, our model could be used to reconstruct garments originally designed by third parties, potentially raising issues of copyright and intellectual property infringement. We emphasize that our framework is intended for research and responsible use, and any deployment in commercial settings should ensure compliance with applicable copyright laws and respect for designer rights.

Reproducibility Statement

This work uses only public datasets and open-source models for its training and evaluations. In the Appendix, we include all the implementation and dataset details to reproduce our results. In addition, we will publicly release the source code and trained models to further support reproducibility.

References

Omri Avrahami, Or Patashnik, Ohad Fried, Egor Nemchinov, Kfir Aberman, Dani Lischinski, and Daniel Cohen-Or. Stable flow: Vital layers for training-free image editing. In *CVPR*, 2025.

Shuai Bai, Huiling Zhou, Zhikang Li, Chang Zhou, and Hongxia Yang. Single Stage Virtual Try-On Via Deformable Attention Flows. In *ECCV*, 2022.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025.

Alberto Baldrati, Davide Morelli, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Multimodal Garment Designer: Human-Centric Latent Diffusion Models for Fashion Image Editing. In *ICCV*, 2023.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. *arXiv preprint arXiv:1801.01401*, 2018.

Chieh-Yun Chen, Yi-Chung Chen, Hong-Han Shuai, and Wen-Huang Cheng. Size Does Matter: Size-aware Virtual Try-on via Clothing-oriented Transformation Try-on Network. In *ICCV*, 2023.

Zhennan Chen, Yajie Li, Haofan Wang, Zhibo Chen, Zhengkai Jiang, Jun Li, Qian Wang, Jian Yang, and Ying Tai. Region-aware text-to-image generation via hard binding and soft refinement. In *ICCV*, 2024.

Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. VITON-HD: High-Resolution Virtual Try-On via Misalignment-Aware Normalization. In *CVPR*, 2021.

Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for authentic virtual try-on in the wild. In *ECCV*, 2024.

Zheng Chong, Xiao Dong, Haoxiang Li, Shiyue Zhang, Wenqing Zhang, Xujie Zhang, Hanqing Zhao, Dongmei Jiang, and Xiaodan Liang. CatVTON: Concatenation Is All You Need for Virtual Try-On with Diffusion Models. In *ICLR*, 2025.

Aiyu Cui, Daniel McKee, and Svetlana Lazebnik. Dressing in Order: Recurrent Person Image Generation for Pose Transfer, Virtual Try-On and Outfit Editing. In *ICCV*, 2021.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image Quality Assessment: Unifying Structure and Texture Similarity. *IEEE Trans. PAMI*, 44(5):2567–2581, 2020.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In *ICML*, 2024.

Benjamin Fele, Ajda Lampe, Peter Peer, and Vitomir Struc. C-VTON: Context-Driven Image-Based Virtual Try-On Network. In *WACV*, 2022.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data. In *NeurIPS*, 2023.

Daniel Garibi, Shahar Yadin, Roni Paiss, Omer Tov, Shiran Zada, Ariel Ephrat, Tomer Michaeli, Inbar Mosseri, and Tali Dekel. TokenVerse: Versatile Multi-concept Personalization in Token Modulation Space. In *SIGGRAPH*, 2025.

Federico Girella, Davide Talon, Ziyue Liu, Zanxi Ruan, Yiming Wang, and Marco Cristani. LOTS of Fashion! Multi-Conditioning for Image Generation via Sketch-Text Pairing. In *ICCV*, 2025.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. *Communications of the ACM*, 63(11):139–144, 2020.

Hailong Guo, Bohan Zeng, Yiren Song, Wentao Zhang, Chuang Zhang, and Jiaming Liu. Any2AnyTryon: Leveraging Adaptive Position Embeddings for Versatile Virtual Clothing Tasks. *arXiv preprint arXiv:2501.15891*, 2025.

Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. VITON: An Image-Based Virtual Try-On Network. In *CVPR*, 2018.

Alec Helbling, Tuna Han Salih Meral, Ben Hoover, Pinar Yanardag, and Duen Horng Chau. Conceptattention: Diffusion transformers learn highly interpretable features. In *ICML*, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*, 2022.

Xun Huang and Serge Belongie. Arbitrary Style Transfer in Real-Time With Adaptive Instance Normalization. In *ICCV*, 2017.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

Boyuan Jiang, Xiaobin Hu, Donghao Luo, Qingdong He, Chengming Xu, Jinlong Peng, Jiangning Zhang, Chengjie Wang, Yunsheng Wu, and Yanwei Fu. FitDiT: Advancing the Authentic Garment Details for High-fidelity Virtual Try-on. In *CVPR*, 2025.

Jeongho Kim, Gyojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. StableVITON: Learning Semantic Correspondence with Latent Diffusion Model for Virtual Try-On. In *CVPR*, 2024.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-Resolution Virtual Try-On with Misalignment and Occlusion-Handled Conditions. In *ECCV*, 2022.

Seungyong Lee and Jeong-gi Kwak. Voost: A Unified and Scalable Diffusion Transformer for Bidirectional Virtual Try-On and Try-Off. *arXiv preprint arXiv:2508.04825*, 2025.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *CVPR*, 2023.

Han Lin, Jaemin Cho, Abhay Zala, and Mohit Bansal. Ctrl-adapter: An efficient and versatile framework for adapting diverse controls to any diffusion model. In *ICLR*, 2025.

Kuan Heng Lin, Sicheng Mo, Ben Klingher, Fangzhou Mu, and Bolei Zhou. Ctrl-x: Controlling structure and appearance for text-to-image generation without guidance. In *NeurIPS*, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. In *ICLR*, 2023.

Jinxi Liu, Zijian He, Guangrun Wang, Guanbin Li, and Liang Lin. One Model For All: Partial Diffusion for Unified Try-On and Try-Off in Any Pose. *arXiv preprint arXiv:2508.04559*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICLR*, 2019.

Davide Morelli, Fincato Matteo, Cornia Marcella, Landi Federico, Cesari Fabio, and Cucchiara Rita. Dress Code: High-Resolution Multi-Category Virtual Try-On. In *ECCV*, 2022.

Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. LaDI-VTON: Latent Diffusion Textual-Inversion Enhanced Virtual Try-On. In *ACM Multimedia*, 2023.

Chong Mou et al. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, 2024.

Weili Nie, Sifei Liu, Morteza Mardani, Chao Liu, Benjamin Eckart, and Arash Vahdat. Compositional text-to-image generation with dense blob representations. In *ICML*, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning Robust Visual Features without Supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On Aliased Resizing and Surprising Subtleties in GAN Evaluation. In *CVPR*, 2022.

William Peebles and Saining Xie. Scalable Diffusion Models with Transformers. In *ICCV*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

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

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: Memory Optimizations Toward Training Trillion Parameter Models. In *SC*, 2021.

Bin Ren, Hao Tang, Fanyang Meng, Ding Runwei, Philip HS Torr, and Nicu Sebe. Cloth Interactive Transformer for Virtual Try-On. *ACM TOMM*, 20(4):1–20, 2023a.

Bin Ren, Hao Tang, Yiming Wang, Xia Li, Wei Wang, and Nicu Sebe. PI-Trans: Parallel-convmlp and implicit-transformation based Gan for cross-view image translation. In *ICASSP*, 2023b.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *NeurIPS*, 2019.

Ivan Skorokhodov, Sharath Girish, Benran Hu, Willi Menapace, Yanyu Li, Rameen Abdal, Sergey Tulyakov, and Aliaksandr Siarohin. Improving the diffusability of autoencoders. In *ICML*, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. In *ICLR*, 2021.

Nick Stracke, Stefan Andreas Baumann, Kolja Bauer, Frank Fundel, and Björn Ommer. Cleandift: Diffusion features without noise. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 117–127, 2025.

Yanan Sun, Yanchen Liu, Yinhao Tang, Wenjie Pei, and Kai Chen. Anycontrol: create your artwork with versatile control on text-to-image generation. In *ECCV*, 2024.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *ICCV*, 2025.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. In *CVPR*, 2023.

Riza Velioglu, Petra Bevandic, Robin Chan, and Barbara Hammer. TryOffDiff: Virtual-Try-Off via High-Fidelity Garment Reconstruction using Diffusion Models. In *BMVC*, 2024.

Riza Velioglu, Petra Bevandic, Robin Chan, and Barbara Hammer. MGT: Extending Virtual Try-Off to Multi-Garment Scenarios. In *ICCV Workshops*, 2025.

Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward Characteristic-Preserving Image-based Virtual Try-On Network. In *ECCV*, 2018.

Haoxuan Wang, Jinlong Peng, Qingdong He, Hao Yang, Ying Jin, Jiafu Wu, Xiaobin Hu, Yanjie Pan, Zhenye Gan, Mingmin Chi, et al. Unicombine: Unified multi-conditional combination with diffusion transformer. In *ICCV*, 2025.

Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation. In *CVPR*, 2024.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Processing*, 13(4):600–612, 2004.

Ioannis Xarchakos and Theodoros Koukopoulos. TryOffAnyone: Tiled Cloth Generation from a Dressed Person. *arXiv preprint arXiv:2412.08573*, 2024.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *CVPR*, 2025.

Zhenyu Xie, Zaiyu Huang, Xin Dong, Fuwei Zhao, Haoye Dong, Xijin Zhang, Feida Zhu, and Xiaodan Liang. GP-VTON: Towards General Purpose Virtual Try-on via Collaborative Local-Flow Global-Parsing Learning. In *CVPR*, 2023.

Yuhao Xu, Tao Gu, Weifeng Chen, and Chengcai Chen. OOTDiffusion: Outfitting Fusion based Latent Diffusion for Controllable Virtual Try-on. In *AAAI*, 2025.

Keyu Yan, Tingwei Gao, Hui Zhang, and Chengjun Xie. Linking garment with person via semantically associated landmarks for virtual try-on. In *CVPR*, 2023.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. *arXiv preprint arXiv:2308.06721*, 2023.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation Alignment for Generation: Training Diffusion Transformers Is Easier Than You Think. In *ICLR*, 2025.

Jianhao Zeng, Dan Song, Weizhi Nie, Hongshuo Tian, Tongtong Wang, and An-An Liu. CAT-DM: Controllable Accelerated Virtual Try-on with Diffusion Model. In *CVPR*, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training. In *ICCV*, 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *ICCV*, 2023.

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

Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. In *NeurIPS*, 2023.

Zijian Zhou, Shikun Liu, Xiao Han, Haozhe Liu, Kam Woh Ng, Tian Xie, Yuren Cong, Hang Li, Mengmeng Xu, Juan-Manuel Pérez-Rúa, Aditya Patel, Tao Xiang, Miaojing Shi, and Sen He. Learning Flow Fields in Attention for Controllable Person Image Generation. In *CVPR*, 2025.

Luyang Zhu, Dawei Yang, Tyler Zhu, Fitsum Reda, William Chan, Chitwan Saharia, Mohammad Norouzi, and Ira Kemelmacher-Shlizerman. TryOnDiffusion: A Tale of Two UNets. In *CVPR*, 2023.

Luyang Zhu, Yingwei Li, Nan Liu, Hao Peng, Dawei Yang, and Ira Kemelmacher-Shlizerman. M&m vto: Multi-garment virtual try-on and editing. In *CVPR*, 2024.

A Additional Details

A.1 Datasets Details

Dress Code. In our experiments, we adopt the Dress Code dataset (Morelli et al., 2022), the largest publicly available benchmark for image-based virtual try-on. Unlike previous datasets limited to upper-body clothing, Dress Code includes three macro-categories: upper-body clothes with 15,363 pairs (*e.g.*, tops, t-shirts, shirts, sweatshirts), lower-body clothing with 8,951 pairs (*e.g.*, trousers, skirts, shorts), and full-body dresses with 29,478 pairs. The total number of paired samples is $53,792$, split into $48,392$ training images and $5,400$ test images at a resolution of 1024×768.

VITON-HD. Following previous literature, we also adopt VITON-HD (Choi et al., 2021), a publicly available dataset widely used in virtual try-on research. It is composed exclusively of upper-body garments and provides high-resolution images at 1024×768 pixels. The dataset contains a total of $27,358$ images, structured into $13,679$ garment-model pairs. These are split into $11,647$ training pairs and $2,032$ test pairs, each comprising a front-view image of a garment and the corresponding image of a model wearing it.

A.2 Implementation Details

For both the feature extractor and the diffusion backbone, we adopt Stable Diffusion 3 medium (Esser et al., 2024). All models are trained on a single node equipped with 4 NVIDIA A100 GPUs (64GB each), using DeepSpeed ZeRO-2 (Rajbhandari et al., 2021) for efficient distributed training. We use a total batch size of 32 and train each model for 30k steps, corresponding to approximately 960k images. Optimization is performed with AdamW (Loshchilov & Hutter, 2019), using a learning rate of 1×10^{-4}, a warmup phase of 3k steps, and a cosine annealing schedule. We train separate models per dataset to account for differences in distribution and garment structure. In all experiments, we set the alignment loss weight λ equal to 0.5.

We evaluate our method both with distribution-based metrics and per-sample similarity metrics. For the first group, we adopt FID (Parmar et al., 2022) and KID (Bińkowski et al., 2018) implementations derived from `clean-fid` PyTorch package[2]. Concerning the second group, we adopt both SSIM (Wang et al., 2004) and LPIPS (Zhang et al., 2018) as they are the standard metrics adopted in the field to measure structural and perceptual similarity between a pair of images. We reuse the corresponding Python packages provided by TorchMetrics[3]. Finally, we adopt DISTS (Ding et al., 2020) as an additional sample-based similarity metric, as it correlates better with human judgment, as shown in previous works (Fu et al., 2023). We stick to the corresponding Python package[4] to compute it for our experiments.

A.3 Caption Extraction Details

We leverage Qwen2.5-VL (Bai et al., 2025) to generate the caption of a given garment image, following the chat template provided below:

```
visual_attributes = {
    "dresses": ["Cloth Type", "Waist", "Fit", "Hem", "Neckline", "Sleeve
        Length", "Cloth Length"],
    "upper_body": ["Cloth Type", "Waist", "Fit", "Hem", "Neckline", "
        Sleeve Length", "Cloth Length"],
    "lower_body": ["Cloth Type", "Waist", "Fit", "Cloth Length"]
}
```

System: You are Qwen, created by Alibaba Cloud. You are a helpful assistant.

[2] https://pypi.org/project/clean-fid/
[3] https://pypi.org/project/torchmetrics/
[4] https://pypi.org/project/DISTS-pytorch/

User:

Use only visual attributes that are present in the image. Predict values of the following attributes: `{visual_attributes[category]}`.
It's forbidden to generate the following visual attributes: colors, background, and textures/patterns. It's forbidden to generate unspecified predictions. It's forbidden to generate newline characters. Generate in this way: `a <cloth type> with <attributes description>.`

Qwen Caption:
A denim shirt with a straight fit, long sleeves, and a button-down neckline. The hem is straight and the shirt appears to be of standard length.

We decide to generate structural-only attributes because our base model without text can already transfer colors and textures correctly from the person image to the generated garment image. The structural attributes are slightly different according to the three categories of clothing, as specified in `visual_attributes`. For example, the neckline can be specified for upper body and dresses (whole body garments), but not for lower body items.

A.4 Algorithm

To provide a clear understanding of TEMU-VTOFF, we summarize the core components of our method in Algorithm 1. The pseudo-code outlines the sequential steps involved in training our dual-DiT architecture, including multimodal conditioning, the hybrid attention module, and the garment aligner component.

B Additional Quantitative Results and Analyses

In this section, we report additional quantitative results, the effect of the λ parameter, and the effect of the asynchronous timestep conditioning between our F_E and F_D.

Effect of Varying λ Parameter. We conducted an ablation study on the Dress Code dataset to assess the effect of the λ regularization for the alignment of our main diffusion transformer F_D and the DINOv2 features. We report the results in Table 4. As shown, $\lambda = 0.5$ is the overall best choice across all metrics.

Analysis of Asynchronous Timestep Conditioning. A critical design choice in our architecture is the use of a fixed timestep $t = 0$ for the feature extractor F_E, while the main denoising DiT F_D operates on a noisy latent z_t at timestep $t > 0$. This raises an important question: could this discrepancy in timesteps lead to a misalignment in the feature space? In this section, we provide the rationale for this design choice, supported by concurrent work and a targeted ablation study.

Our primary motivation is to provide the main generator F_D with the cleanest, most semantically rich conditioning signal possible. By extracting features from F_E at $t = 0$ we ensure the conditioning information is completely free from stochastic noise inherent to the diffusion process. We hypothesize

Algorithm 1 TEMU-VTOFF: Dual-DiT and Garment Alignment for VTOFF

Require: Person image $\boldsymbol{x}_{\text{model}}$, garment caption c, binary mask M, target garment image $\boldsymbol{x}_g$
Ensure: Generated garment $\hat{\boldsymbol{x}}_g$

Latent encoding:
  Encode the target garment: $\boldsymbol{z}_g \leftarrow \mathcal{E}(\boldsymbol{x}_g)$
  Sample noise: $\epsilon_t \sim \mathcal{N}(0, 1)$
  Apply flow-matching: $\boldsymbol{z}_t \leftarrow (1-t)\boldsymbol{z}_g + t \cdot \epsilon_t$
Prepare masked spatial input:
  Encode masked person image: $\boldsymbol{x}_M \leftarrow \mathcal{E}(\boldsymbol{x}_{\text{model}} \odot M)$
  Concatenate inputs: $\boldsymbol{z}'_t \leftarrow [\boldsymbol{z}_t, M, \boldsymbol{x}_M]$
Extract modulation features:
  $\boldsymbol{e}^v_{\text{pool}} \leftarrow \text{CLIP}(\boldsymbol{x}_{\text{model}})$
Extract keys and values using feature extractor:
  $\{\boldsymbol{K}^l_{\text{extractor}}, \boldsymbol{V}^l_{\text{extractor}}\}_{l=1}^N \leftarrow F_E(\boldsymbol{z}'_0, \boldsymbol{x}_{\text{model}}, t{=}0)$
Encode text information:
  Get pooled text embedding: $\boldsymbol{e}_{\text{pooled}} \leftarrow \text{CLIP}(c)$
  Get full sequence text features: $\boldsymbol{e}_{\text{text}} \leftarrow [\text{CLIP}(c), \text{T5}(c)]$
Noise prediction:
  $\hat{\epsilon}_t \leftarrow F_D(\boldsymbol{z}_t, \boldsymbol{e}_{\text{pooled}}, \boldsymbol{e}_{\text{text}}, \{\boldsymbol{K}^l_{\text{extractor}}, \boldsymbol{V}^l_{\text{extractor}}\}, t)$
  Compute diffusion loss: $\mathcal{L}_{\text{DiT}} \leftarrow \|\hat{\epsilon}_t - \epsilon_t\|^2$
Align internal representations:
  Extract DiT features: $\boldsymbol{h}_{\text{DiT}} \leftarrow$ tokens from 8th block of F_D
  Extract DINOv2 features: $\boldsymbol{h}_{\text{enc}} \leftarrow \text{DINOv2}(\boldsymbol{x}_g)$
  Align via projection: $\tilde{\boldsymbol{h}}_{\text{DiT}} \leftarrow \phi_{\text{CNN}}(\boldsymbol{h}_{\text{DiT}})$
  Compute alignment loss: $\mathcal{L}_{\text{align}} \leftarrow -\frac{1}{N}\sum_i \cos(\tilde{\boldsymbol{h}}^{\text{DiT}}_i, \boldsymbol{h}^{\text{enc}}_i)$
Final objective:
  Combine losses: $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{DiT}} + \lambda \cdot \mathcal{L}_{\text{align}}$
Decode final garment:
  Run reverse process: $\hat{\boldsymbol{x}}_g \leftarrow \mathcal{D}(\hat{\boldsymbol{z}}_0)$

that injecting features from a noisy timestep $t > 0$ would introduce an additional, confounding source of noise into the generation process, thereby degrading the quality of the final output. The key to our method is that the MHA module is specifically trained to bridge this temporal gap; it learns to effectively attend to the clean conditioning features to guide the denoising of the noisy latent z_t.

This design philosophy is strongly supported by recent, concurrent research that analyzes the internal representations of diffusion models:

- The work on CleanDIFT (Stracke et al., 2025) directly argues that adding noise to images before feature extraction is a performance bottleneck that harms feature quality. Their entire method is built on the same premise as our F_E: that extracting features from clean images leads to superior performance without needing task-specific timestep tuning.
- Furthermore, ConceptAttention (Helbling et al., 2025) demonstrates that the internal representations of DiTs are highly interpretable and correspond to semantic concepts, particularly at early timesteps. This validates our choice to use $t = 0$ features, as they represent the purest and most semantically meaningful form of this information.

To validate our design choice, we conducted an ablation study comparing our method with the variant where the feature extractor F_E and the denoising F_D use the same synchronous timestep t. The results on Dress Code are presented in Table 4. As shown in the table, our proposed method with asynchronous timesteps significantly outperforms the synchronous variant across the majority of the metrics. This result provides strong empirical evidence for the value of clean conditioning and confirms the effectiveness of our proposed Multimodal Hybrid Attention.

C Additional Qualitative Results

We report an extended version of the qualitative results presented in our main paper. Specifically, additional visual comparisons between TEMU-VTOFF and competitors are shown in Fig. 6 and Fig. 7, on sample images from Dress Code (Morelli et al., 2022) and VITON-HD (Choi et al., 2021),

Table 4: Additional ablation study results on the Dress Code dataset.

	All					Upper-body		Lower-body		Dresses	
	SSIM ↑	**LPIPS ↓**	**DISTS ↓**	**FID ↓**	**KID ↓**	**DISTS ↓**	**FID ↓**	**DISTS ↓**	**FID ↓**	**DISTS ↓**	**FID ↓**
Effect of Varying λ Parameter											
$\lambda = 0.0$ (w/o garment aligner)	**76.01**	**30.84**	20.63	5.91	0.78	21.77	11.26	22.26	14.22	17.86	11.86
$\lambda = 0.25$	74.29	33.68	19.41	6.42	0.89	**19.65**	**10.77**	21.86	16.53	16.73	11.38
$\lambda = 0.75$	71.93	37.03	20.35	7.81	1.41	20.11	11.09	23.85	21.12	17.10	11.69
$\lambda = 1.0$	71.76	37.21	20.45	7.78	1.39	20.07	11.33	24.11	20.59	17.17	11.79
$\lambda = 0.5$ ***(Ours)***	75.95	31.46	**18.66**	**5.74**	**0.65**	19.75	10.94	**19.57**	**13.83**	**16.67**	**11.29**
Effect of Asynchronous Timestep Conditioning											
w/ same t in F_E and F_D	**77.70**	32.69	22.41	9.78	2.30	23.98	17.85	21.29	17.83	21.95	17.52
w/ $t = 0$ in F_E ***(Ours)***	75.95	**31.46**	**18.66**	**5.74**	**0.65**	**19.75**	**10.94**	**19.57**	**13.83**	**16.67**	**11.29**

respectively. Moreover, Fig. 8 presents additional ablation results to analyze the impact of textual and mask conditioning. Finally, we include in Fig. 9 the full set of inputs used for generating the target garment, including the model input, the segmentation mask, and the textual caption.

D LIMITATIONS

Our model inherits some inner problems of foundational models like Stable Diffusion 3 (Esser et al., 2024). Even though our method improved the generation of big logos and text, it is limited in its scope concerning fine-grained details like complex texture patterns and small written text. Moreover, it sometimes fails to render the correct number of small objects like buttons. We show a set of failure cases in Fig. 10 and Fig. 11, on sample images from Dress Code (Cui et al., 2021) and VITON-HD (Lee et al., 2022) respectively.

E LLM USAGE

In this work, we employ LLMs (specifically Qwen2.5-VL) to extract garment-related textual descriptions, which serve as conditioning signals for generation. Beyond this, LLMs were employed solely for minor language refinement. They did not contribute to the design of experiments, the analysis of results, or the generation of scientific content.

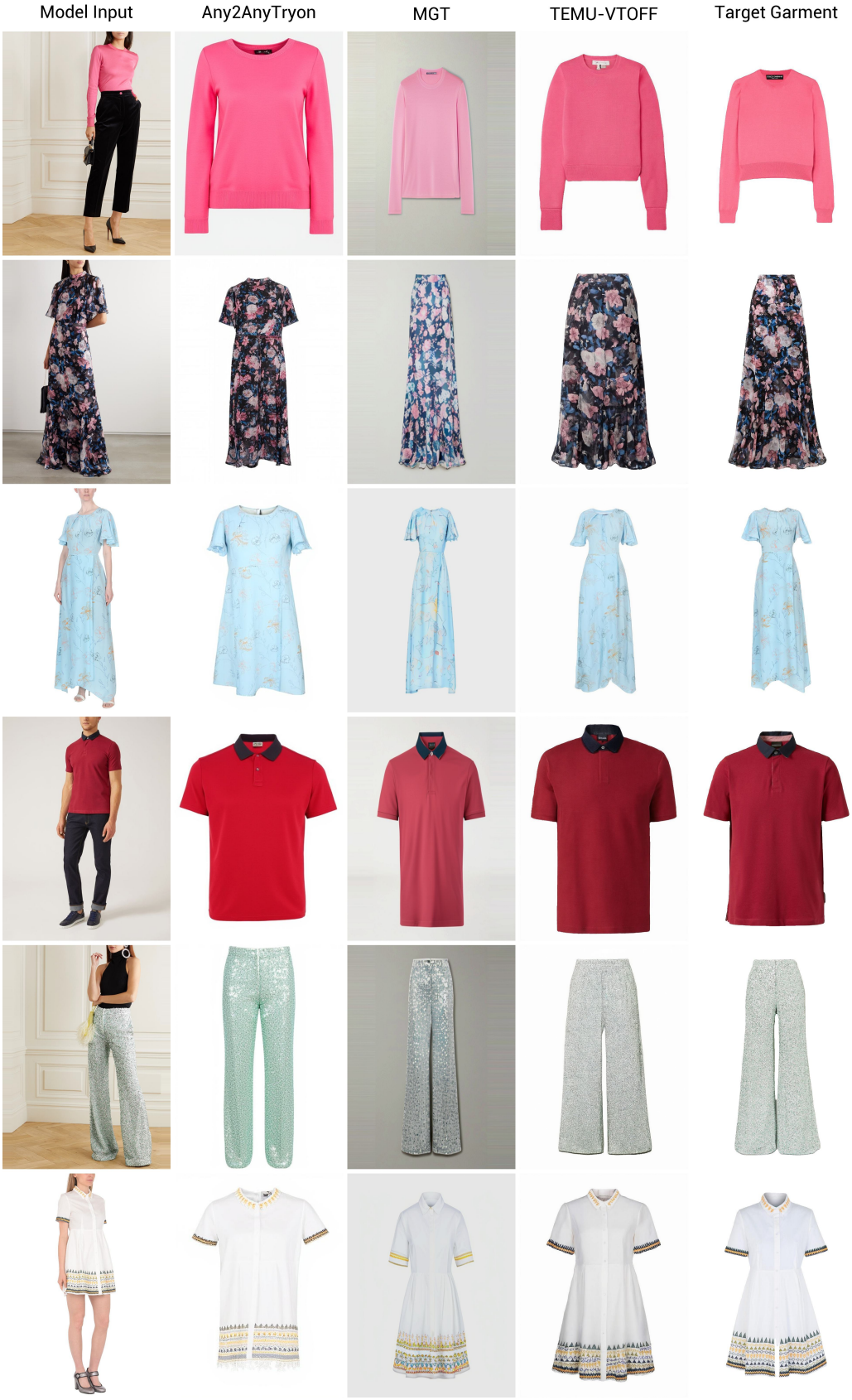


Figure 6: Additional qualitative results of TEMU-VTOFF and competitors on Dress Code (Morelli et al., 2022).

Figure 7: Additional qualitative results of TEMU-VTOFF and competitors on VITON-HD (Choi et al., 2021).

Figure 8: Additional qualitative results showing the contribution of each component in TEMU-VTOFF on Dress Code (Morelli et al., 2022) images.

Figure 9: Inputs used to generate the target garment with TEMU-VTOFF, using sample images from Dress Code (Cui et al., 2021)

Figure 10: An overview of failure cases on the Dress Code (Morelli et al., 2022) dataset.

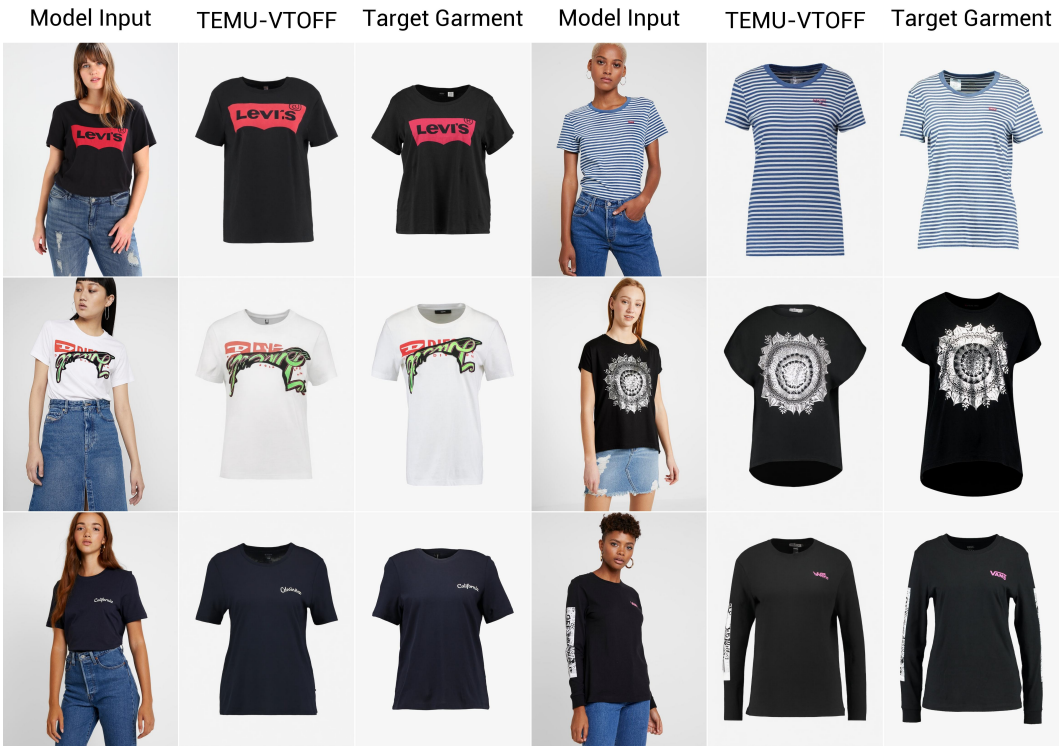


Figure 11: An overview of failure cases on the VITON-HD (Choi et al., 2021) dataset.