TOWO-VTON: TAKE-OFF WEAR-ON VIRTUAL TRY-ON VIA TOPOLOGY-AWARE DECOUPLING

Anonymous authors
Paper under double-blind review



Figure 1: **Effect of ToWo-VTON under large garment transformations.** Large length and category changes, such as long pants to shorts (left) or a long asymmetrical dress to a short polo dress (right), often cause severe topology artifacts without ToWo. *w/ ToWo* denotes Wear-On trained on $\mathcal{D}_{\text{BFDC}}$ and inferred with the two-stage Take-Off/Wear-On pipeline.

ABSTRACT

Virtual Try-On (VTON) aims to synthesize garment transfer without physical trials. Although recent methods perform well on upper-body garments, they often fail in lower-body and full-body scenarios due to complex occlusions and inadequate modeling of body structure. To address this limitation, we introduce ToWo (Take-Off Wear-On), a plug-in two-stage paradigm that decouples body topology reconstruction from garment synthesis. Take-Off first recovers an occlusion-free body representation, and Wear-On then synthesizes the target garment conditioned on this structural prior. This decoupling improves garment placement and robustness, especially in challenging lower-body and full-body settings with large silhouette variations and cross-category transformations. Built on this paradigm, we propose ToWo-VTON, a unified topology-aware framework with structure-aware supervision. We construct a new Body-Fitting Dress-Code (BFDC) dataset via body-aware diffusion and VLM-based filtering to provide topology-consistent supervision across upper-, lower-, and full-body cases, particularly benefiting scenarios with severe occlusions. To demonstrate generality, we instantiate ToWo-VTON on both diffusion-based and MLLM-guided backbones. Diffusion models adopt a plug-in conditioning strategy with BFDC fine-tuning, while MLLM-guided models incorporate Multimodal Alignment Blocks (MABs) for semantic–topology alignment and Garment Boosting Attention (GBA) for detail enhancement. Extensive experiments demonstrate that ToWo-VTON achieves state-of-the-art performance, reducing KID by 26.1% and 51.5% across diffusion-based and MLLM-guided backbones, respectively. It also produces more coherent and perceptually realistic results across diverse garment categories.

1 INTRODUCTION

Virtual Try-On (VTON) has emerged as a promising technology in fashion and e-commerce, enabling realistic garment previews without requiring physical trials (Han et al., 2018; Kim et al., 2024; Choi et al., 2025; Chong et al., 2025; Yang et al., 2024; Kim et al., 2025). Early methods

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

relied on Generative Adversarial Networks (GANs) (Goodfellow et al., 2020), which often suffered from limited generative capacity, leading to low visual fidelity and restricted garment diversity. More recently, diffusion-based generation (Radford et al., 2018) has improved realism and structural alignment, achieving strong performance in upper-body synthesis where occlusions are limited and geometric correspondence is easier to model.

However, despite their strengths, existing diffusion-based VTON methods struggle to generalize to lower-body and full-body try-on scenarios. These tasks require an understanding of more complex spatial relationships between garments and the human body, especially under occlusions and pose diversity. Without explicit modeling of body–garment topology, current systems frequently produce artifacts such as distorted limbs, misaligned clothing, and inconsistent garment textures. In addition, these issues are exacerbated in attempts to scale beyond upper-body clothing categories, highlighting a key limitation in current VTON systems.

To address this challenge, we introduce ToWo (Take-Off Wear-On), a unified two-stage paradigm that explicitly decouples body topology reconstruction from garment synthesis. Take-Off first removes the original garment and reconstructs an occlusion-reduced, topology-consistent body representation, effectively eliminating garment-induced structural interference. Wear-On then synthesizes the target garment conditioned on this clean structural prior. By removing clothing occlusions before generation, ToWo prevents the model from relying on entangled garment cues to infer body geometry, enabling more stable and coherent synthesis under large garment coverage and severe occlusions. Importantly, this formulation is compatible with existing diffusion-based VTON architectures and naturally extends to multimodal large language model (MLLM) backbones.

To further provide structure-aware supervision, we construct the Body-Fitting Dress-Code (BFDC) dataset via a self-refinement strategy. We bootstrap a body-aware diffusion model on a curated tight-fit subset and filter generated samples using a vision-language model to ensure structural consistency. The resulting dataset supplies topology-consistent structural priors that strengthen geometry-aware alignment during training, particularly under challenging occlusion and silhouette variations.

To incorporate the topology awareness into garment synthesis, we design a unified Wear-On framework for both diffusion-based and MLLM-guided generative backbones. By replacing the original person image, whose body structure is entangled with the source garment, with the body structure proxy, our pipeline removes source-garment interference and provides clearer body and limb topology for target garment generation. For diffusion-based VTON backbones, our pipeline directly replaces the original person input with the body proxy and fine-tunes the model on BFDC without architectural modification. For the MLLM-guided backbone, Wear-On uses Multimodal Alignment Blocks (MABs) to align the body proxy with garment-related multimodal conditions before denoising, while Garment Boosting Attention (GBA) fuses garment features to improve texture fidelity and garment-body coherence. Together, these designs support both plug-in diffusion adaptation and MLLM-guided generation through the same topology-preserving conditioning framework.

Extensive evaluations demonstrate that ToWo-VTON consistently outperforms prior SOTA methods, with particularly pronounced gains in geometrically challenging settings. Our approach reduces KID by 26.1% and 51.5% on DressCode (Morelli et al., 2022) with diffusion-based and MLLM-guided backbones, respectively, highlighting the effectiveness of topology decoupling across architectures. User studies further confirm substantial improvements in structural consistency and perceptual realism across diverse garment categories. Our main contributions are summarized as follows:

- We identify garment-body topology entanglement as a core limitation of existing VTON systems and introduce ToWo, a two-stage Take-Off/Wear-On paradigm that decouples body topology acquisition from garment synthesis.
- We develop the Body Structure Revealing (BSR) model and construct the Body-Fitting Dress-Code (BFDC) dataset, providing topology-consistent body proxies for both inference-time conditioning and training-time supervision.
- We propose a unified Wear-On framework that supports plug-in adaptation to diffusion-based VTON backbones and further scales to an MLLM-guided DiT architecture with MABs and GBA for semantic alignment and garment-detail enhancement.
- We validate ToWo-VTON across upper-body, lower-body, and full-body try-on, achieving improvements over state-of-the-art baselines in quantitative metrics and human evaluations.

2 Related Works

GAN-based VTON. Early VTON methods relied on GANs (Goodfellow et al., 2020) with appearance flow and garment warping (Xie et al., 2023; He et al., 2022; Chopra et al., 2021; Yu et al., 2019; Dong et al., 2019; Ge et al., 2021; Chen et al., 2021; Wang et al., 2018). VITON (Han et al., 2018) introduced pose-guided content warping for garment alignment, followed by condition refinement to better estimate body–garment correspondence (Cui et al., 2021; Yang et al., 2020). Later works like SD-VITON (Shim et al., 2024) incorporated pose maps for improved alignment. S-WUTON (Issenhuth et al., 2020) introduced a teacher–student paradigm for garment transfer. However, GAN-based methods often suffer from limited generalization and visual artifacts.

Diffusion-based VTON. Diffusion models have recently enabled more flexible and realistic garment transfer through denoising-based generation (Yang et al., 2024; Liang et al., 2025; Hong et al., 2025). Unlike warping-based methods, they leverage image inpainting to integrate garments more naturally (Wan et al., 2025a; Yan et al., 2023; Li et al., 2023; Kim et al., 2025; Zhang et al., 2025a). Works such as StableVITON (Kim et al., 2024) and IDM-VTON (Choi et al., 2025) improve identity preservation via structured conditioning, but often require additional inputs (e.g., pose maps or 3D cues (He et al., 2025)) that introduce computational overhead and noise sensitivity. To reduce reliance on manual inputs, recent methods adopt data-driven pipelines with enhanced identity and text control (Chong et al., 2025; Du et al., 2025; Zhang et al., 2025b; Wan et al., 2025b; Ye et al., 2023). Yet, most works focus on upper-body garments, with lower-body VTON remaining challenging due to garment diversity (e.g., skirts, dresses) and occlusion handling (Chen et al., 2025; Yang et al., 2025; Zhu et al., 2024; Li et al., 2025; Guo et al., 2025; Chen et al., 2023).

MLLMs for Image Editing. MLLMs support vision-language tasks, including VTON, by aligning visual and textual features (Dong et al., 2024). Models like GenArtist (Wang et al., 2024) and VITRON (Fei et al., 2024) enable fine-grained control via language prompts, while Qwen-Image (Wu et al., 2025a) and DreamOmni2 (Xia et al., 2025) offer strong cross-modal representations. However, current MLLM approaches often fail to capture structural relationships between garments and bodies. To address this, we incorporate MLLM supervision to enhance contextual reasoning and improve body–garment alignment across diverse clothing categories.

3 Methodology

3.1 Body Topology Acquisition via Take-Off

Existing single-stage virtual try-on (VTON) systems often conflate body shape with garment appearance. This entanglement becomes problematic when clothing is loose or occludes key body regions, leading to inaccurate body reconstruction and misaligned garment placement, especially in cross-category try-on tasks with varying silhouettes. To address this issue, we introduce Take-Off, the first stage of ToWo, which recovers a garment-neutral structural prior before synthesis. Take-Off approximates an occlusion-reduced body by replacing the original garment with a tight-fitting template that preserves geometric cues such as limb contours, joint positions, and pose structure, serving as a proxy for body topology without requiring explicit body exposure.

BSR model. Take-Off is operationalized through the Body Structure Revealing (BSR) model. Given a person image with potentially loose or occluding clothing, BSR infers a clean, pose-consistent body proxy by translating the original appearance into a tight-fitting surrogate that reveals the underlying body structure. This surrogate is not intended to reconstruct private body appearance, but rather to expose the geometric cues required for topology-aware garment synthesis.

To obtain BSR, we reproduce BooW-VTON (Zhang et al., 2025c) and fine-tune it for tight-fit body surrogate reconstruction. Specifically, we construct a compact training set $\mathcal{D}_{\text{BF}}$ from the DressCode dataset (Morelli et al., 2022), containing N paired samples $(X_m^{\text{fit}}, X_g^{\text{fit}})$. Here, each $g_i \in \mathcal{G}^{\text{fit}} = \{g_1, \ldots, g_{10}\}$ denotes a manually selected tight-fitting garment template, designed to reveal underlying body shape and articulation across diverse poses and body types. The selected tight-fit templates are visualized in Appendix E.

As illustrated in Fig. 2(a) Stage I, we synthesize occluded inputs by re-dressing each tight-fitting person image X_m^{fit} with a randomly sampled garment X_g and its textual description D_g using our repro-

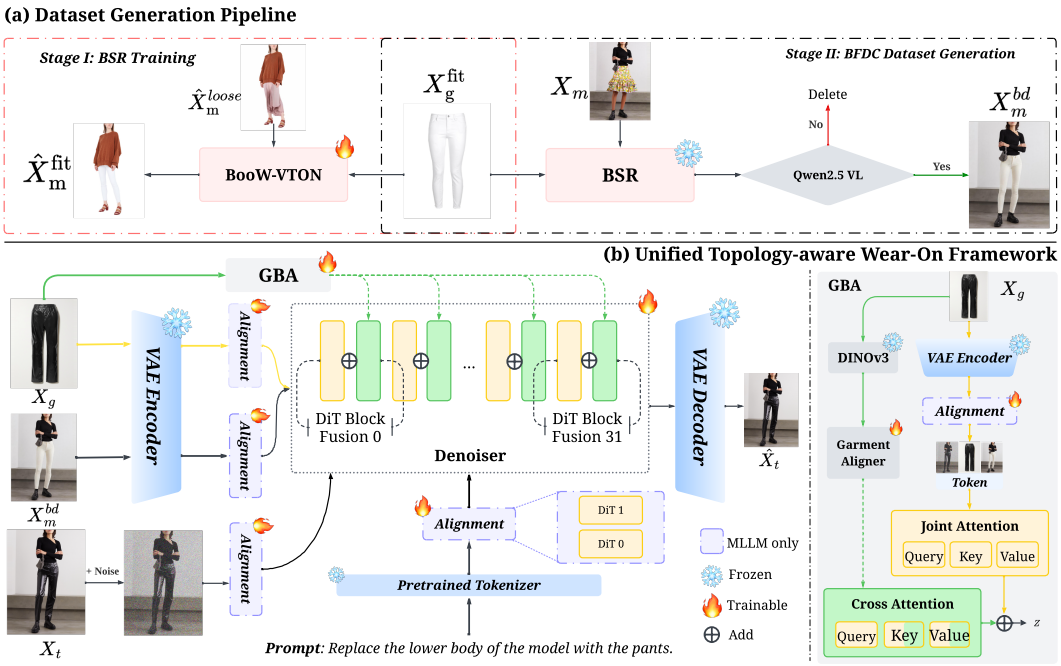

Figure 2: **Overview of the ToWo-VTON framework.** (a) *Dataset Generation*: Stage I uses reproduced BooW-VTON to create loose-clothing variants from tight-fit samples, enabling fine-tuning into the Body Structure Revealing (BSR) model. Stage II applies BSR to generate tight-fit surrogates for DressCode images, forming the BFDC dataset after filtering by Qwen2.5-VL. (b) *Unified Topology-aware Wear-On framework*: Wear-On uses the Take-Off surrogate X_m^{bd} as the body-conditioning input for garment synthesis. Existing diffusion-based VTON backbones can be adapted by replacing the garment-entangled person image X_m with X_m^{bd}, while the MLLM-based variant strengthens semantic-topology alignment and detail preservation through MABs and GBA.

duced BooW-VTON. This produces a loose-clothing variant $\hat{X}_m^{\text{loose}}$ that simulates realistic garment-induced occlusion while preserving the underlying subject identity and pose. We then form training triplets $(X_g^{\text{fit}}, X_m^{\text{fit}}, \hat{X}_m^{\text{loose}})$, where X_g^{fit} serves as the tight-fit template condition, X_m^{fit} is the desired reconstruction target, and $\hat{X}_m^{\text{loose}}$ is the occluded input to be restored.

Using these triplets, we fine-tune the reproduced BooW-VTON into the BSR model, which reconstructs the tight-fitting reference from the occluded input conditioned on the tight-fit template:

$$\hat{X}_m^{\text{fit}} = \text{BSR}(\hat{X}_m^{\text{loose}} \mid X_g^{\text{fit}}, D_g^{\text{fit}}). \tag{1}$$

The model is optimized using the original diffusion training objective, encouraging BSR to suppress garment-induced geometric ambiguity while preserving identity, pose, and body structure.

Body-Fitting DressCode (BFDC) Dataset. While Take-Off provides a body proxy decoupled from the source garment at inference time, effective Wear-On synthesis also requires the generator to be trained with the same type of structural conditioning. If the generator is trained only on garment-entangled person images but tested with Take-Off surrogates, the distribution gap may weaken the benefit of topology-aware conditioning. Therefore, we construct Body-Fitting DressCode (BFDC), a geometry-aware training set that extends structural decoupling from inference to training.

As illustrated in Fig. 2(a) Stage II, BFDC is constructed strictly from the official DressCode training split (Morelli et al., 2022), containing 8,951 lower-body, 29,478 dress, and 15,363 upper-body pairs. The 1,800 test pairs in each category are excluded before proxy generation. For each training person image X_m, we apply the trained BSR model using a selected tight-fit template $(X_g^{\text{fit}}, D_g^{\text{fit}})$:

$$X_m^{\text{bd}} = \text{BSR}(X_m \mid X_g^{\text{fit}}, D_g^{\text{fit}}). \tag{2}$$

The resulting surrogate X_m^{bd}, generated at the original DressCode resolution of 768×1024, preserves identity and pose while replacing the source garment with a tight-fitting structural proxy. We form training triplets $(X_m, X_g, X_m^{\text{bd}})$, where X_m^{bd} provides geometry-aware conditioning for Wear-On.

To ensure data quality, we utilize Qwen2.5-VL (Qwen et al., 2025) as a filter during BFDC construction. For each candidate surrogate, the VLM performs artifact inspection based on valid and invalid exemplars with brief rationales, including loose-fitting garments, disconnected or duplicated limbs, and overlapping clothing artifacts. For lower-body samples, we apply a tight-fit-to-legs criterion to ensure that leg topology is clearly visible. The VLM outputs a binary decision, and only verified samples are retained. The dataset, denoted as $\mathcal{D}_{\text{BFDC}}$, provides topology-consistent supervision for upper-body, lower-body, and full-body try-on. Take-Off and $\mathcal{D}_{\text{BFDC}}$ establish a unified structural decoupling paradigm: Take-Off supplies clean body proxies at inference time, while $\mathcal{D}_{\text{BFDC}}$ aligns the Wear-On generator with the same decoupled conditioning distribution during training.

3.2 Unified Topology-aware Wear-On Framework

In the Wear-On stage, we use X_m^{bd} to denote the body proxy produced by Take-Off, either from $\mathcal{D}_{\text{BFDC}}$ during training or from the input person image during inference. Wear-On synthesizes the target try-on image using X_m^{bd} as structural conditioning. Compared with the original person image X_m, which entangles body structure with source-garment appearance, X_m^{bd} provides a cleaner topology carrier, including pose, body contour, limb configuration, and spatial layout. By replacing X_m with X_m^{bd}, Wear-On passes the recovered topology into garment synthesis, while X_g and the garment-related prompt provide appearance and semantic guidance.

Given a garment-related prompt P, a frozen pretrained encoder produces semantic tokens:

$$T = \mathcal{P}_{\text{enc}}(P), \tag{3}$$

where $\mathcal{P}_{\text{enc}}$ denotes the pretrained encoder. For diffusion-based VTON backbones, $\mathcal{P}_{\text{enc}}$ follows the original CLIP/T5 encoders. For our MLLM-based design, it is instantiated with Qwen2.5-VL to provide stronger vision-grounded semantic guidance.

At an abstract level, Wear-On performs conditional denoising with three complementary sources of information: semantic guidance from T, body topology from X_m^{bd}, and garment appearance from X_g. This generic conditioning interface can be written as:

$$\hat{\epsilon} = \mathcal{D}_\theta\left(\mathbf{z}_t, t \mid T, \mathcal{E}(X_m^{\text{bd}}), \mathcal{E}(X_g)\right), \tag{4}$$

where $\mathcal{D}_\theta$ denotes the diffusion denoiser and $\mathcal{E}$ denotes a pretrained VAE encoder. This abstraction defines a unified conditioning interface that covers both plug-in adaptation to existing diffusion-based VTON backbones and our MLLM-guided DiT implementation, where these conditioning signals are further aligned and refined before denoising, as shown in Fig. 2(b).

3.2.1 Plug-in Adaptation to Diffusion-based Backbones

We instantiate this plug-in design on two representative diffusion-based VTON backbones: the mask-based TED-VITON (Wan et al., 2025c) and the mask-free BooW-VTON (Zhang et al., 2025c). Both backbones already contain a person-conditioning branch, allowing the Take-Off body proxy to be used as the structural input without changing the network architecture, diffusion objective, or sampling pipeline. Given a diffusion-based VTON generator $\mathcal{F}$, the adapted process is:

$$\hat{X}_t = \mathcal{F}(X_m^{\text{bd}} \mid X_g, D_g), \tag{5}$$

where X_g provides target garment appearance, D_g provides semantic guidance, and X_m^{bd} provides decoupled body topology. We further fine-tune each backbone on $\mathcal{D}_{\text{BFDC}}$ to align the training distribution with the Take-Off conditioning used at inference time. This enables the generator to interpret X_m^{bd} as its structural input while reducing reliance on source-garment silhouettes.

3.2.2 Scaling to MLLM-based Backbones

Although diffusion-based adaptation improves topology-aware conditioning, its semantic reasoning and cross-modal alignment remain constrained by the original VTON backbone. We therefore scale Wear-On to an MLLM-based DiT architecture, where the topology from X_m^{bd}, the appearance from X_g, and the semantic tokens from Qwen2.5-VL are jointly modeled during generation. In this design, Wear-On follows a topology-preserving workflow: MABs first align the body proxy with garment-related conditions before denoising, the DiT denoiser then propagates the aligned topology during generation, and GBA fuses DINOv3 garment features into DiT layers for detail preservation.

Table 1: Quantitative results on the DressCode Lower-body and Dresses datasets. Ours (TED) and Ours (BooW) are topology-aware plug-in variants, while Ours is the MLLM-based. w/o ToWo denotes Wear-On without both $\mathcal{D}_{\text{BFDC}}$ training and Take-Off inference. **Bold** and underline indicate the best and second-best results. FID is scaled by 100. * indicates reproduced results.

Dataset	Dress-Code Lower-body				Dress-Code Dresses			
Method	KID↓	FID↓	SSIM↑	LPIPS↓	KID↓	FID↓	SSIM↑	LPIPS↓
Image Editing Methods								
QWEN-Image (Wu et al., 2025a)	0.907	19.438	0.842	0.212	0.389	15.667	0.826	0.189
DreamOmni2 (Xia et al., 2025)	0.848	21.669	0.801	0.286	1.900	35.636	0.738	0.301
GAN-based VTON Methods								
SD-VITON (Shim et al., 2024)	6.202	69.156	0.825	0.370	8.830	85.921	0.801	0.378
Diffusion-based VTON Methods								
StableVITON (Kim et al., 2024)	0.966	19.732	0.878	0.216	0.707	23.454	0.850	0.210
IDM-VTON (Choi et al., 2025)	0.497	16.125	0.871	0.193	0.359	14.065	0.860	0.173
CAT-VTON (Chong et al., 2025)	1.012	20.146	0.875	0.232	0.640	17.210	0.841	0.211
TED-VITON* (Wan et al., 2025c)	0.762	19.500	0.877	0.184	0.497	15.032	0.846	0.182
BooW-VTON* (Zhang et al., 2025c)	0.509	16.724	0.854	0.210	0.583	17.526	0.844	0.172
Ours (TED)	0.584	14.663	0.890	0.183	0.408	14.926	**0.865**	0.168
Ours (BooW)	0.376	13.952	0.892	0.172	0.491	15.849	0.858	0.166
MLLM-based VTON Methods								
Ours w/o ToWo	0.415	14.201	0.877	0.179	1.026	23.007	0.826	0.200
Ours	**0.172**	**11.044**	**0.895**	**0.154**	**0.190**	**12.459**	0.860	**0.152**

Multimodal Alignment Blocks (MABs). To align heterogeneous conditioning signals before iterative denoising, we introduce Multimodal Alignment Blocks (MABs). Each MAB jointly processes four modalities: the semantic token T, the body latent $\mathcal{E}(X_m^{\text{bd}})$, the garment latent $\mathcal{E}(X_g)$, and Gaussian noise $\mathbf{z}$ sampled from the diffusion prior. Architecturally, each MAB consists of two consecutive DiT blocks with joint self-attention, applied before the main DiT denoiser to refine these heterogeneous inputs into an aligned latent space for subsequent generation:

$$(T_r, X_m^r, X_g^r, \mathbf{z}_r) = \text{MAB}(T, \mathcal{E}(X_m^{\text{bd}}), \mathcal{E}(X_g), \mathbf{z}), \tag{6}$$

where $\text{MAB}(\cdot)$ denotes the two-layer DiT and the superscript r denotes the refined representations. By performing joint attention over semantic, body, garment, and noise tokens before denoising, MABs align garment semantics with body topology. This reduces cross-modal inconsistency and provides the downstream DiT denoiser with coherent and semantically grounded representations.

Topology-conditioned Denoiser. After MABs, the refined representations $(T_r, X_m^r, X_g^r, \mathbf{z}_r)$ are fed into the main DiT denoiser for iterative generation. The aligned body representation X_m^r carries the topology recovered from the Take-Off proxy, while T_r and X_g^r provide semantic guidance and target garment appearance. Conditioned on these aligned representations, the denoiser progressively updates $\mathbf{z}_r$ to synthesize the try-on result. In this process, X_m^r remains in the main generation stream as a topology-consistent structural condition, guiding the global pose, body contour, and garment–body spatial layout across diffusion steps. Thus, Wear-On generates the target garment based on the recovered wearer topology rather than the garment-entangled source image.

Garment Boosting Attention (GBA). Although MABs align the tokens before denoising, fine-grained garment details may still be weakened during generation. We therefore introduce Garment Boosting Attention (GBA), which extracts garment features from DINOv3 (Siméoni et al., 2025) and maps them into garment-specific keys and values through a learnable Garment Aligner G:

$$(K_g, V_g) = G(\text{DINOv3}(X_g)). \tag{7}$$

These features are fused into each DiT layer via cross-attention:

$$Z = \text{Attention}(Q, K_j, V_j) + \lambda\, \text{Attention}(Q, K_g, V_g), \tag{8}$$

where Q, K_j, and V_j are from the main denoising stream, and λ controls the strength of garment feature fusion. By injecting DINOv3 garment features into the topology-conditioned denoising stream, GBA provides each DiT layer with fine-grained garment cues, enhancing texture fidelity, preserving local details, and improving garment-body coherence under the recovered body topology.



Figure 3: Qualitative comparison on (a) DressCode Lower-body and (b) DressCode Dresses. Ours (TED) and Ours (BooW) are topology-aware diffusion variants, while Ours is the full MLLM-based ToWo-VTON. Across both garment categories, our models exhibit more consistent garment–body alignment and clearer garment details than competing approaches. The improvements are reflected in more coherent garment placement and better preservation of fine-grained visual characteristics, yielding try-on results that maintain both structural consistency and garment fidelity.

4 Experiments

4.1 Experiment Setup

Datasets. We evaluate our model on two widely used VTON benchmarks: VITON-HD (Choi et al., 2021) and DressCode (Morelli et al., 2022). Both datasets are augmented using the mask-free BooW-VTON (Zhang et al., 2025c) framework to establish a standardized and reproducible benchmark and enable fair evaluation under unpaired settings across garment categories. We adopt the official training and test splits. For Dress-Code, 1,800 pairs are selected for testing from each subset: 8,951 lower-body, 29,478 dress, and 15,363 upper-body pairs. For VITON-HD, 2,032 of 13,679 pairs are used for testing, with the remaining samples utilized for training. All images are processed at a resolution of 768 $\times$ 1024 to ensure consistency across models and datasets.

Baselines. We compare ToWo-VTON with representative methods across four categories. For image editing approaches, we include recent multimodal generative systems QWEN-Image (Wu et al., 2025a) and DreamOmni2 (Xia et al., 2025). For GAN-based VTON, we adopt SD-VITON (Shim et al., 2024). Diffusion-based VTON baselines include StableVITON (Kim et al., 2024), IDM-VTON (Choi et al., 2025), CAT-VTON (Chong et al., 2025), TED-VITON (Wan et al., 2025c), and BooW-VTON (Zhang et al., 2025c). For fair comparison, we reproduce TED-VITON and BooW-VTON under a unified evaluation protocol. To evaluate the plug-in capability of our topology-aware framework, we apply Take-Off to two representative diffusion paradigms and fine-tune them on $\mathcal{D}_{\text{BFDC}}$, namely the mask-based TED-VITON and the mask-free BooW-VTON, denoted as Ours (TED) and Ours (BooW), respectively. These variants retain the original architectures, training objectives, and diffusion schedules, differing only in the proposed topology-aware conditioning.

To the best of our knowledge, prior work has not explored MLLM-guided VTON. We therefore construct a direct MLLM baseline by adapting OmniGen2 (Wu et al., 2025b) to the try-on task, de-

Table 2: Quantitative comparison on VITON-HD and DressCode Upper-body benchmarks. Ours (TED) and Ours (BooW) are topology-aware plug-in variants, while Ours is the MLLM-based. w/o ToWo denotes Wear-On without both $\mathcal{D}_{\text{BFDC}}$ training and Take-Off inference. **Bold** and underline indicate the best and second-best results. FID is scaled by 100. * indicates reproduced results.

Dataset	VITON-HD				Dress-Code Upper-body			
Method	KID↓	FID↓	SSIM↑	LPIPS↓	KID↓	FID↓	SSIM↑	LPIPS↓
Image Editing Methods								
QWEN-Image (Wu et al., 2025a)	0.794	16.499	0.688	0.381	0.920	19.112	0.890	0.223
DreamOmni2 (Xia et al., 2025)	0.551	15.076	0.743	0.345	1.528	30.054	0.771	0.325
GAN-based VTON Methods								
SD-VITON (Shim et al., 2024)	0.280	10.552	0.864	0.204	6.143	67.833	0.847	0.335
Diffusion-based VTON Methods								
StableVITON (Kim et al., 2024)	0.575	14.105	0.843	0.232	0.920	19.112	0.890	0.222
IDM-VTON (Choi et al., 2025)	0.423	12.248	0.839	0.225	0.304	11.821	0.915	0.160
CAT-VTON (Chong et al., 2025)	0.554	12.847	0.828	0.266	0.645	14.550	0.897	0.208
TED-VITON* (Wan et al., 2025c)	0.359	12.063	0.845	0.221	0.373	13.480	0.909	0.178
BooW-VTON* (Zhang et al., 2025c)	0.254	11.621	0.830	0.234	0.289	10.642	0.915	0.170
Ours (TED)	0.184	8.968	0.863	0.200	0.278	12.107	**0.922**	0.186
Ours (BooW)	0.194	9.868	0.847	0.231	0.220	**9.815**	0.918	0.157
MLLM-based VTON Methods								
Ours w/o ToWo	0.225	10.028	0.826	0.246	0.296	13.834	0.902	0.189
Ours	**0.176**	**8.673**	**0.868**	**0.195**	**0.194**	10.532	0.920	**0.149**

noted as Ours w/o ToWo, which uses neither Take-Off inference nor fine-tuning on $\mathcal{D}_{\text{BFDC}}$. Our full model, denoted as Ours, integrates topology-aware structural decoupling within the same backbone for comparison. All models are evaluated at a unified resolution of 768×1024, with outputs rescaled when necessary to ensure fairness. A detailed analysis of the computational overhead introduced by the Take-Off stage and the MLLM-based Wear-On implementation is provided in Appendix G.

Evaluation Metrics. To evaluate the performance of VTON approaches in real-world settings, we compare model performance across upper-body, lower-body, and full-body categories, where the models are tasked with reconstructing the ground-truth garments from humans wearing altered garments, ensuring fair metric evaluation. We adopt four widely used VTON metrics: Fréchet Inception Distance (FID) (Heusel et al., 2018) and Kernel Inception Distance (KID) (Bińkowski et al., 2021) to assess dataset-level distributional alignment and realism, Structural Similarity Index (SSIM) (Wang et al., 2004) to evaluate structural fidelity, and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) to quantify abstract visual feature differences. Fairness of the unpaired evaluation protocol is provided in Appendix H. To complement these automatic metrics, we further conduct a user study in Appendix F to evaluate perceptual realism and garment fitting quality.

4.2 Qualitative Results

As shown in Fig. 3, the advantages of our topology-aware design are most evident in lower-body and dress synthesis, where garment topology frequently changes in length, looseness, and silhouette. Existing methods tend to overfit to the original garment structure, resulting in incorrect waist alignment, distorted leg regions, or residual source-clothing patterns. In contrast, our method decouples body topology from garment appearance, enabling better inference of human structure and yielding more natural length transitions, improved waist and hip alignment, and clearer textures. The improvement becomes more pronounced in full-body dress synthesis (Fig. 3b), where the garment must be coherently generated across the entire body. Prior approaches often struggle with global consistency, producing discontinuities or structural distortions when large portions of the body are covered. Our topology-aware framework maintains smoother global structure and more coherent garment geometry, leading to more realistic dress fitting and better preservation of garment details. Additional upper-body and in-the-wild results are provided in Appendix D and L. We further verify the contribution of the topology-aware conditioning strategy through ablation studies in Appendix C.

4.3 Quantitative Results

Tables 1 and 2 present quantitative comparisons on VITON-HD and DressCode across upper-, lower-, and full-body settings. Overall, ToWo-VTON consistently outperforms prior methods, with the most notable improvements observed in lower-body and dress scenarios. These tasks involve larger garment topology changes, such as variations in garment length, looseness, and silhouette. As shown



Figure 4: **Topology-aware attention analysis.** ToWo shifts diffusion attention from garment regions to underlying human structural areas, demonstrating improved topology-aware conditioning. *w/o ToWo* denotes Wear-On trained without $\mathcal{D}_{\text{BFDC}}$ and inferred directly from the original input.

in Table 1, our model achieves the best performance on the DressCode Lower-body benchmark and ranks best or second-best across all metrics on DressCode Dresses, demonstrating clear and consistent improvements in garment–body alignment under significant geometric variations.

For diffusion-based VTON, we apply Take-Off during inference and fine-tune the mask-based TED-VITON and mask-free BooW-VTON on $\mathcal{D}_{\text{BFDC}}$, leading to consistent gains. The resulting variants, Ours (TED) and Ours (BooW), outperform their original counterparts in KID and FID while maintaining competitive perceptual quality. In the MLLM setting, a direct baseline without structural decoupling (Ours w/o ToWo) performs noticeably worse. After introducing topology-aware conditioning, the full model achieves the best overall performance across most benchmarks, confirming the effectiveness of topology-aware structural decoupling for realistic virtual try-on.

4.4 Topology Awareness Analysis

To further validate that the proposed ToWo mechanism enables topology-aware reasoning, we visualize the attention maps of the diffusion backbone under different conditioning inputs, as shown in Fig. 4. Specifically, when the original person image is directly used as input (w/o ToWo), the model's attention tends to concentrate around residual garment regions and source-clothing boundaries, reflecting the entanglement between clothing appearance and body structure. In contrast, after replacing the input with the topology-neutral ToWo body representation, the attention shifts toward semantically meaningful human structural areas, such as the waistline, hips, and leg contours, which are critical for estimating garment placement and body–garment alignment. This redistribution of attention indicates that the body proxy generated by ToWo suppresses garment-induced geometric bias and encourages the model to focus more consistently on underlying human topology. The results provide feature-level evidence that our structural decoupling strategy enhances topology-aware representation learning during generation, especially under challenging clothing occlusions.

5 Conclusions

In this paper, we present ToWo-VTON, a topology-aware virtual try-on framework that decouples body structure reconstruction from garment synthesis. Our two-stage paradigm recovers a garment-neutral body representation via Take-Off and then performs garment synthesis through Wear-On conditioning. We construct the Body-Fitting DressCode (BFDC) dataset to provide topology-consistent supervision across diverse garment categories. Extensive experiments on VITON-HD and DressCode demonstrate that ToWo-VTON achieves state-of-the-art performance with improved garment–body alignment and visual realism under challenging pose and silhouette variations. Failure cases and limitations are discussed in Appendix I.

AI USE STATEMENT

For manuscript preparation, we used generative AI tools only for language polishing, including grammar correction and improvements to clarity and readability. We did not use AI assistance for research ideation, methodology or experimental design, code development, or interpretation of results. As described in the paper, our research pipeline uses generative models to construct body-fitting proxy images for $\mathcal{D}_{\text{BFDC}}$ and vision-language models for proxy filtering and failure detection. We reviewed all AI-assisted textual revisions to ensure that they preserved the intended technical meaning. We take full responsibility for the final content of this work, including its text, claims, results, and artifacts.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we describe the two-stage ToWo framework and the construction and filtering procedures of $\mathcal{D}_{\text{BFDC}}$ in the main text. Implementation details are provided in Section A, including model initialization, training datasets, Wear-On backbone configurations, optimization settings, and computational hardware.

REFERENCES

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans, 2021. URL https://arxiv.org/abs/1801.01401.

Chieh-Yun Chen, Ling Lo, Pin-Jui Huang, Hong-Han Shuai, and Wen-Huang Cheng. Fashionmirror: Co-attention feature-remapping virtual try-on with sequential template poses. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13809–13818, October 2021.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7513–7522, October 2023.

Mengting Chen, Xi Chen, Zhonghua Zhai, Chen Ju, Xuewen Hong, Jinsong Lan, and Shuai Xiao. Wear-any-way: Manipulable virtual try-on via sparse correspondence alignment. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 124–142, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72630-9.

Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14131–14140, June 2021.

Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for authentic virtual try-on in the wild. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 206–235, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73016-0.

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, February 2025. URL http://arxiv.org/abs/2407.15886. arXiv:2407.15886 [cs].

Ayush Chopra, Rishabh Jain, Mayur Hemani, and Balaji Krishnamurthy. Zflow: Gated appearance flow-based virtual try-on with 3d priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5433–5442, October 2021.

Aiyu Cui, Daniel McKee, and Svetlana Lazebnik. Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14638–14647, October 2021.

Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, and Jian Yin. Towards multi-pose guided virtual try-on network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning, 2024. URL https://arxiv.org/abs/2301.00234.

Chenghu Du, Shengwu Xiong, and Yi Rong. All parts matter: A unified mask-free virtual try-on framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19525–19534, October 2025.

Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 57207–57239. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/68bad5506f0f9eea7ae75f01ae00d5e2-Paper-Conference.pdf.

Chongjian Ge, Yibing Song, Yuying Ge, Han Yang, Wei Liu, and Ping Luo. Disentangled cycle consistency for highly-realistic virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16928–16937, June 2021.

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, Jiaming Liu, and Chuang Zhang. Any2anytryon: Leveraging adaptive position embeddings for versatile virtual clothing tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19085–19096, October 2025.

Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S. Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Sen He, Yi-Zhe Song, and Tao Xiang. Style-based global appearance flow for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3470–3479, June 2022.

Zijian He, Yuwei Ning, Yipeng Qin, Guangrun Wang, Sibei Yang, Liang Lin, and Guanbin Li. Vton 360: High-fidelity virtual try-on from any viewing direction. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 26388–26398, June 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018. URL https://arxiv.org/abs/1706.08500.

Ji Woo Hong, Tri Ton, Trung X. Pham, Gwanhyeong Koo, Sunjae Yoon, and Chang D. Yoo. Itamdt: Image-timestep-adaptive masked diffusion transformer framework for image-based virtual try-on. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 28284–28294, June 2025.

Thibaut Issenhuth, Jérémie Mary, and Clément Calauzènes. Do not mask what you do not need to mask: A parser-free virtual try-on. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 619–635, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58565-5.

Jeongho Kim, Guojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8176–8185, June 2024.

Jeongho Kim, Hoiyeong Jin, Sunghyun Park, and Jaegul Choo. Promptdresser: Improving the quality and controllability of virtual try-on via generative textual prompt and prompt-aware mask. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16026–16036, October 2025.

Black Forest Labs. Flux.1-fill. https://github.com/black-forest-labs/flux, 2024. Accessed: 2025-11-05.

Xinghui Li, Qichao Sun, Pengze Zhang, Fulong Ye, Zhichao Liao, Wanquan Feng, Songtao Zhao, and Qian He. Anydressing: Customizable multi-garment virtual dressing via latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 23723–23733, June 2025.

Zhi Li, Pengfei Wei, Xiang Yin, Zejun Ma, and Alex C. Kot. Virtual try-on with pose-garment keypoints guided inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22788–22797, October 2023.

Yujie Liang, Xiaobin Hu, Boyuan Jiang, Donghao Luo, Xu Peng, Kai Wu, Chengming Xu, Wenhui Han, Taisong Jin, Chengjie Wang, and Rongrong Ji. Vton-handfit: Virtual try-on for arbitrary hand pose guided by hand priors embedding. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 22616–22626, June 2025.

Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: High-resolution multi-category virtual try-on. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 345–362, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-20074-8.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 Technical Report, January 2025. URL http://arxiv.org/abs/2412.15115. arXiv:2412.15115 [cs].

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *OpenAI Blog*, 2018. URL https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

Sang-Heon Shim, Jiwoo Chung, and Jae-Pil Heo. Towards squeezing-averse virtual try-on via sequential deformation. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i5.28288. URL https://doi.org/10.1609/aaai.v38i5.28288.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. URL https://arxiv.org/abs/2508.10104.

Siqi Wan, Jingwen Chen, Yingwei Pan, Ting Yao, and Tao Mei. Incorporating Visual Correspondence into Diffusion Model for Virtual Try-On, May 2025a. URL http://arxiv.org/abs/2505.16977. arXiv:2505.16977 [cs].

Zhenchen Wan, Yanwu Xu, Dongting Hu, Weilun Cheng, Tianxi Chen, Zhaoqing Wang, Feng Liu, Tongliang Liu, and Mingming Gong. Mft-viton: High-fidelity virtual try-on with minimal input via a mask-free transformer-diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 1985–1994, October 2025b.

Zhenchen Wan, Yanwu Xu, Zhaoqing Wang, Feng Liu, Tongliang Liu, and Mingming Gong. Tedviton: Transformer-empowered diffusion models for virtual try-on, 2025c. URL https://arxiv.org/abs/2411.17017.

Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *The European Conference on Computer Vision (ECCV)*, September 2018.

Shansong Wang, Mingzhe Hu, Qiang Li, Mojtaba Safari, and Xiaofeng Yang. Capabilities of gpt-5 on multimodal medical reasoning, 2025. URL https://arxiv.org/abs/2508.08224.

Zhenyu Wang, Aoxue Li, Zhenguo Li, and Xihui Liu. Genartist: Multimodal llm as an agent for unified image generation and editing. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 128374–128395. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/e7c786024ca718f2487712bfe9f51030-Paper-Conference.pdf.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025a. URL https://arxiv.org/abs/2508.02324.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. Omnigen2: Exploration to advanced multimodal generation, 2025b. URL https://arxiv.org/abs/2506.18871.

Bin Xia, Bohao Peng, Yuechen Zhang, Junjia Huang, Jiyang Liu, Jingyao Li, Haoru Tan, Sitong Wu, Chengyao Wang, Yitong Wang, Xinglong Wu, Bei Yu, and Jiaya Jia. Dreamomni2: Multimodal instruction-based editing and generation, 2025. URL https://arxiv.org/abs/2510.06679.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23550–23559, June 2023.

Keyu Yan, Tingwei Gao, Hui Zhang, and Chengjun Xie. Linking garment with person via semantically associated landmarks for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17194–17204, June 2023.

Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Xu Yang, Changxing Ding, Zhibin Hong, Junhao Huang, Jin Tao, and Xiangmin Xu. Texture-preserving diffusion models for high-fidelity virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7017–7026, June 2024.

Zhaotong Yang, Yuhui Li, Shengfeng He, Xinzhe Li, Yangyang Xu, Junyu Dong, and Yong Du. Omnivton: Training-free universal virtual try-on. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16702–16711, October 2025.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2308.06721.

Ruiyun Yu, Xiaoqi Wang, and Xiaohui Xie. Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Delong Zhang, Qiwei Huang, Yang Sun, Yuanliu Liu, Wei-Shi Zheng, Pengfei Xiong, and Wei Zhang. Learning implicit features with flow-infused transformations for realistic virtual try-on. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18736–18745, October 2025a.

Nannan Zhang, Yijiang Li, Dong Du, Zheng Chong, Zhengwentai Sun, Jianhao Zeng, Yusheng Dai, Zhengyu Xie, Hairui Zhu, and Xiaoguang Han. Robust-mvton: Learning cross-pose feature alignment and fusion for robust multi-view virtual try-on. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 16029–16039, June 2025b.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018. URL https://arxiv.org/abs/1801.03924.

Xuanpu Zhang, Dan Song, Pengxin Zhan, Tianyu Chang, Jianhao Zeng, Qingguo Chen, Weihua Luo, and An-An Liu. Boow-vton: Boosting in-the-wild virtual try-on via mask-free pseudo data training. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 26399–26408, June 2025c.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1346–1356, June 2024.

ToWo-VTON: Take-Off Wear-On Full-Body Virtual Try-On via Topology-aware Decoupling

Supplementary Material

A Implementation Details.

We follow the two-stage ToWo paradigm. The Body Structure Revealing (BSR) model is initialized from BooW-VTON (Zhang et al., 2025c) and fine-tuned on the BF dataset using AdamW with a learning rate of 1.0×10^{-5} and $(\beta_1, \beta_2) = (0.9, 0.999)$. For the Wear-On stage, we fine-tune three backbones, namely TED-VITON, BooW-VTON, and an attention-modified OmniGen2, on the $\mathcal{D}_{\text{BFDC}}$ dataset with topology-consistent conditioning. TED-VITON and BooW-VTON follow their original training learning rates, while the OmniGen2 variant is optimized with AdamW using $(\beta_1, \beta_2) = (0.9, 0.95)$ and a learning rate of 1.0×10^{-6} for 10 epochs. All experiments are conducted on a single NVIDIA H100 GPU.

B Upper–Lower Difficulty Analysis

Existing VTON methods exhibit a consistent performance gap between upper-body and lower-body synthesis, with lower-body results showing significantly higher KID values. This reflects the increased difficulty of modeling complex occlusions, pose variations, and garment–body interactions in lower-body scenarios. Table 3 reports the KID scores for upper- and lower-body synthesis on DressCode. We observe that baseline methods consistently present a substantial upper-lower gap. After integrating Take-Off and BFDC, this gap is reduced across multiple backbones. Notably, for our model, the KID gap decreases from 0.119 (w/o ToWo) to 0.022, indicating improved consistency between upper- and lower-body synthesis. This result demonstrates that topology-aware structural decoupling effectively mitigates the intrinsic difficulty disparity between garment categories.

Table 3: **Upper-Lower KID gap on DressCode.** ToWo consistently reduces the performance gap between lower- and upper-body synthesis across backbones.

	Lower-body KID↓	Upper-body KID↓	Gap
IDM-VTON (Choi et al., 2025)	0.497	0.304	0.193
CAT-VTON (Chong et al., 2025)	1.012	0.645	0.367
TED-VITON (Wan et al., 2025c)	0.762	0.373	0.389
+ToWo	0.584	0.278	0.306
BooW-VTON (Zhang et al., 2025c)	0.509	0.289	0.220
+ToWo	0.376	0.220	0.156
Ours (w/o ToWo)	0.415	0.296	0.119
Ours	**0.172**	**0.194**	**0.022**

C Ablation Studies of MLLM-based ToWo-VTON

Table 4: Ablation study on the effect of GBA and ToWo. **Bold** indicates the best result. FID is multiplied by 100.

Method	KID↓	FID↓	SSIM↑	LPIPS↓
w/o ToWo	0.415	14.201	0.877	0.179
w/o GBA	0.586	17.775	0.857	0.201
Ours	**0.172**	**11.044**	**0.895**	**0.154**

Effect of ToWo. To assess the effectiveness of ToWo, we apply the same training strategy to enhance two representative diffusion-based methods, TED-VITON (Wan et al., 2025c) and BooW-VTON (Zhang et al., 2025c), where each model is individually fine-tuned using BFDC-augmented data. As shown in Table 4, removing the ToWo module (w/o ToWo) leads to noticeable degradation across all metrics, demonstrating that topology-aware conditioning plays a critical role in improving



(a) Effect of ToWo. **(b)** Effect of GBA.

Figure 5: **Effects of ToWo and GBA.** (a) The first column shows the original input and target garment, while the second shows the occlusion-reduced person image produced by ToWo. (b) The first two columns show the input garment and person. Zoom in for finer details.

structural consistency and perceptual quality. In contrast, the full model achieves the best performance across all metrics. Visual comparisons on the DressCode lower-body dataset (Fig. 5(a)) further reveal improved recovery of limb structures and cleaner garment boundaries, successfully transforming occluding pants into well-aligned shorts without artifacts around the legs.

Effect of Garment Boosting Attention (GBA). To assess the impact of the GBA module, we analyze its effect on fine-grained garment details and garment-body separation. As shown in Fig. 5(b), removing GBA leads to missing or distorted garment elements, such as misaligned buttons and imprecise pant shapes. With GBA, our model accurately reproduces detailed patterns while better distinguishing between the body and the original clothing in the input. This results in clearer structural separation and more faithful garment reconstruction. Quantitatively, as reported in Table 4, incorporating GBA significantly improves perceptual and structural quality, reducing LPIPS from 0.201 to 0.154 and increasing SSIM from 0.857 to 0.895. These improvements confirm that GBA strengthens the model's ability to retain garment identity and spatial coherence during generation.

D In-the-Wild (ITW) Try-On

Prior VTON methods often require explicit In-the-Wild adaptation to handle diverse and complex backgrounds. In contrast, our model achieves robust In-the-Wild (ITW) try-on without requiring specialized background handling. This is made possible through the synergistic interaction between MABs and GBA. MABs reason over multimodal signals to more effectively identify which regions correspond to body topology or garment placement, while GBA enhances garment-specific features and spatial coherence. Together, they allow the model to distinguish between foreground and background regions, ensuring that background areas remain visually untouched during synthesis. As a result, our method produces accurate and consistent garment transformations under diverse and challenging backgrounds, without requiring domain-specific adaptation.

D.1 Dress-Code ITW dataset curation

Inspired by recent approaches using inpainting models for data augmentation (Wan et al., 2025b), we expand the in-the-wild domain coverage by leveraging the FLUX.1-Fill (Labs, 2024) inpainting model to replace the backgrounds of human images in the Dress-Code Lower-Body (Morelli et al., 2022) dataset with diverse and realistic environments. We generate contextual background prompts P_{context} using GPT-5 (Wang et al., 2025), which describe in detail common real-world scenarios for virtual try-on applications. Formally, the inpainting process is defined as:

$$X_{m_bg} = \text{FLUX.1-Fill}\big(X_{m_raw}, P_{\text{context}}\big), \tag{9}$$

where X_{m_raw} denotes the background-removed human image obtained by erasing the background regions guided by the accompanying semantic map X_{sm} provided in Dress-Code. This enables

a rigorous evaluation of model robustness to unseen environments and background interference, quantifying the reliability and realism of virtual try-on synthesis.

Table 5: Quantitative results on the Dress-Code Lower-Body In-The-Wild dataset. **Bold** and underline indicate the best and second-best results. FID is scaled by 100. Results marked with an asterisk (*) denote reproduced implementations.

Dataset	Dress-Code ITW			
Method	KID↓	FID↓	SSIM↑	LPIPS↓
Image Editing Methods				
QWEN-Image (Wu et al., 2025a)	2.012	32.440	0.422	0.532
DreamOmni2 (Xia et al., 2025)	1.173	26.264	0.519	0.412
VTON Methods				
SD-VITON (Shim et al., 2024)	8.060	87.053	0.581	0.458
StableVITON (Kim et al., 2024)	7.517	82.616	0.489	0.595
IDM-VTON (Choi et al., 2025)	0.562	17.838	0.689	0.292
CAT-VTON (Chong et al., 2025)	0.760	20.376	0.675	0.313
TED-VITON* (Wan et al., 2025c)	0.960	24.587	0.707	0.302
BooW-VTON* (Zhang et al., 2025c)	0.525	17.512	0.712	0.297
Ours	**0.333**	**16.046**	**0.722**	**0.285**

D.2 Ablation on ITW Try-On

As shown in Table 5, our method achieves significant improvements in image realism, evidenced by up to a 37% enhancement in KID and an 8.4% improvement in FID compared to existing SOTA approaches, despite not being specifically trained on the new ITW dataset. This performance gain can be attributed to our reasoning blocks, which effectively distinguish between changeable and static regions during garment replacement. Furthermore, our method demonstrates superior structural preservation, as reflected in an improvement of up to $+0.01$ in SSIM and a reduction of -0.007 in LPIPS, indicating enhanced perceptual similarity and structural coherence. Consistent with the observations in Fig. 8, existing VTON methods exhibit intolerance to background variations, while image editing approaches struggle with visual consistency. In contrast, our method demonstrates strong robustness to these challenges, precisely presenting try-on results with intended garment shape and realistic human identities across scenes.

E Tight-Fit Template Selection and Inference Robustness

For Stage 1 (Take-Off), we use manually selected tight-fit garment templates, including 10 upper-body and 10 lower-body garments (Fig. 6). Given the small template set, manual curation is simpler and more reliable than automated VLM-based selection, ensuring clear body topology exposure across diverse and representative poses. During inference, one template is randomly sampled to generate the body proxy X_m^{bd}. Importantly, Stage 2 remains conditioned on the target garment and further incorporates X_m^{bd} only as an attention-fused structural condition rather than hard ground-truth supervision. This design makes the pipeline robust to rare imperfections from Stage 1 generation and effectively avoids over-dependence on any specific strong template prior.



Figure 6: Visualization of Chosen Fit Templates.

F User Study

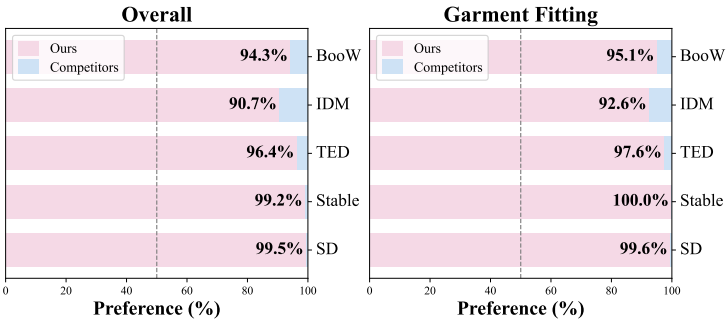

Figure 7: User study comparing our method against each baseline, across overall appearance and garment fitting. Bars show user preference percentages.

To assess perceptual quality, we conducted a user study on lower-body garment try-on results from the DressCode dataset. We selected 40 representative image sets and compared our method (ToWo-VTON) against five baselines: SD-VITON, StableVITON, IDM-VTON, and reproduced TED-VITON, BooW-VTON. Participants were shown side-by-side outputs and asked to choose which result appeared more realistic and visually plausible. A total of 100 valid responses were collected. As shown in Fig. 7, our method was consistently preferred across all pairwise comparisons. In terms of overall appearance, ToWo-VTON received over 90% of user votes against each baseline, reaching 99.5% vs. SD-VITON and 96.4% vs. TED-VITON. For garment fitting, preferences were even stronger. Our model achieved 100% preference over StableVITON and over 95% against most other methods. These results highlight ToWo-VTON's clear advantage in preserving garment structure, shape fidelity, and body–garment alignment, yielding more natural and convincing results.

G Computational Overhead

Table 6 compares the inference efficiency of ToWo-VTON with representative image editing and VTON baselines on an NVIDIA H100, while Table 7 further breaks down the component-wise cost of the MLLM-based Wear-On stage. Since Stage 1 is implemented as a BSR variant fine-tuned from BooW-VTON, it shares the same backbone and therefore has the same inference cost as BooW-VTON. The additional computational cost of ToWo-VTON mainly comes from Stage 2, where the MLLM-based DiT denoiser performs iterative garment synthesis.

Within Stage 2, the Qwen2.5-VL encoder is called only once per inference, while the DiT denoiser dominates the runtime and FLOPs. MABs introduce limited overhead, requiring 4.332s over 50 denoising steps, which is about 6.5% of the main DiT runtime and 3.3% of its FLOPs. Overall, ToWo-VTON is more computationally demanding than single-stage VTON baselines, but remains more efficient than general-purpose MLLM-based image editing methods in both latency and memory. This provides a practical trade-off between efficiency and topology-aware try-on quality.

H Fairness of Unpaired Evaluation

The unpaired evaluation ensures fairness by constructing human-garment triplets from the same person wearing two different garments. The task is to reconstruct the original image by wearing back the unseen garment onto the matching ground truth person image. Even if the initial try-on input person image contains artifacts, the wear-back task itself is applied uniformly. The final metrics are computed identically for all comparing models, aligning with ground-truth human images, to guarantee an unbiased comparison of their garment formation capabilities.

Table 6: Inference efficiency comparison with image editing and VTON baselines. Stage 1 is the BSR variant fine-tuned from BooW-VTON and has the same inference cost as BooW-VTON.

Method	Time (s)	Peak VRAM (GB)	FLOPs (T)
MLLM-Based Methods			
QWEN-Image	546.50	59.96	26121.59
DreamOmni2	312.253	53.97	14367.07
VTON Methods			
IDM-VTON	11.911	13.113	694.09
BooW-VTON	10.359	10.98	1221.8
Ours			
Shared Stage 1	10.359	10.98	1221.8
Stage 2 (Diffusion-based)	10.359	10.98	1221.8
Stage 2 (MLLM-based)	73.789	27.44	31848.8

Table 7: Component-wise cost of the MLLM-based Wear-On stage over 50 denoising steps.

Component	Time (s)	Peak VRAM (GB)	FLOPs (T)
Qwen2.5VL Encoder	2.827	22.56	1.46
MABs (50 steps)	4.332	26.25	1018.88
DiT (50 steps)	66.821	25.96	30823.04

I LIMITATION

Although ToWo-VTON improves topology-aware garment synthesis by decoupling body structure acquisition from garment generation, its performance depends on Take-Off reliability. If Take-Off produces an inaccurate body proxy, such as incorrect body shape or incomplete garment removal, the erroneous structural prior may propagate to Wear-On, causing distorted garment geometry or unrealistic limb structures. As a fallback, Direct Wo bypasses the proxy and performs Wear-On directly from the original input image. Since Wear-On is trained on $\mathcal{D}_{\text{BFDC}}$, it can still learn topology-aware region localization and garment–body correspondence, producing plausible results in some failure cases. However, Direct Wo cannot fully remove garment–body entanglement, and robust Take-Off generation under particularly extreme poses, heavy occlusion, or layered clothing remains future work. To more effectively mitigate error propagation from Take-Off to Wear-On, we also introduce an inference-time detection-and-fallback mechanism, as detailed in Section J.

J MITIGATING CASCADING ERRORS.

To determine when to invoke the above fallback, we introduce an inference-time detection-and-fallback mechanism. Our default detector, Qwen2.5-VL-32B, is also used for filtering during BFDC construction. Before a Take-Off proxy is passed to Wear-On, the detector assesses its validity; samples with proxies flagged as invalid are routed to Direct Wo.

Table 8: Proxy-failure detection results on the official Dress-Code lower-body test split. Detection recall measures the proportion of invalid proxies correctly flagged, while valid-retention rate measures the proportion of valid proxies correctly retained. Both rates are computed using the corresponding class totals in each row.

Detector	Invalid detected	Invalid missed	Detection recall ↑	Valid retained	Valid flagged	Valid-retention rate ↑
Qwen2.5-VL-32B	208	177	54.0%	1,322	93	93.4%
Qwen3.8-27B	104	48	68.4%	1,606	42	97.5%

We evaluate this mechanism on all 1,800 samples in the official Dress-Code lower-body test split. Table 8 summarizes the proxy-failure detection results for Qwen2.5-VL-32B and Qwen3.8-27B. The default detector achieves a failure-detection recall of 54.0% and a valid-retention rate of 93.4%, identifying 208 invalid proxies and retaining 1,322 valid proxies. Qwen3.8-27B achieves a failure-

detection recall of 68.4% and a valid-retention rate of 97.5%, correctly identifying 104 invalid proxies and retaining 1,606 valid proxies. However, the counts imply different valid and invalid proxy totals across the two configurations, limiting direct comparison of their detection performance.

Table 9 reports the resulting try-on performance. With the default Qwen2.5-VL-32B detector, detection and fallback improve KID, SSIM, and LPIPS over the original two-stage inference, while slightly increasing FID from 11.044 to 11.132. With Qwen3.8-27B, detection and fallback improve all four metrics: KID decreases from 0.172 to 0.147, FID decreases from 11.044 to 10.893, SSIM increases from 0.895 to 0.906, and LPIPS decreases from 0.154 to 0.141. An oracle configuration, which uses manually annotated proxy-validity labels to determine whether to invoke Direct Wo, achieves the best performance across all four metrics among the evaluated configurations, with a KID of 0.137, an FID of 10.336, an SSIM of 0.916, and an LPIPS of 0.138. These oracle results suggest that more accurate failure detection could provide additional gains.

Table 9: Effect of detection and fallback on Dress-Code lower-body try-on performance. KID and FID evaluate unpaired generation, while SSIM and LPIPS evaluate paired self-reconstruction. In every fallback configuration, proxies flagged as invalid are bypassed and the corresponding samples use Direct Wo. The oracle uses manually annotated proxy-validity labels for routing.

Inference strategy	$\text{KID}_{\text{UN}} \downarrow$	$\text{FID}_{\text{UN}} \downarrow$	SSIM $\uparrow$	LPIPS $\downarrow$
Original two-stage inference	0.172	11.044	0.895	0.154
Qwen2.5-VL-32B detection + fallback	0.153	11.132	0.898	0.149
Qwen3.8-27B detection + fallback	0.147	10.893	0.906	0.141
Oracle detection + fallback	**0.137**	**10.336**	**0.916**	**0.138**

These results support detection and fallback as a means of mitigating cascading errors in the evaluated setting. Nevertheless, missed failures still expose Wear-On to erroneous structural priors, while false positives may bypass useful proxies. The effectiveness of this mechanism therefore remains constrained by detector reliability and the residual garment–body entanglement in Direct Wo.

K Broader Impacts

ToWo-VTON aims to improve virtual try-on by producing accurate garment-body alignment, especially for lower-body and full-body scenarios. The technology may benefit online fashion applications by reducing the need for physical trials, improving user experience, and supporting garment visualization. At the same time, as with other person-image generation systems, the method may be misused to create misleading synthetic images or to manipulate personal appearance without consent. In addition, virtual try-on systems may reflect biases present in the training data, such as limited coverage of body shapes, poses, clothing styles, or demographic groups. To mitigate these concerns, our experiments are conducted on publicly available VTON datasets, and the proposed Take-Off surrogate is designed to expose geometric structure for garment synthesis rather than reconstruct private body appearance. We encourage responsible deployment with user consent, disclosure of synthetic outputs, and evaluation across users, garments, and body types.

L Additional Qualitative Evaluation

We provide additional qualitative results to further evaluate ToWo-VTON across upper-body, lower-body, and full-body garment categories on DressCode and VITON-HD. As shown in Fig. 8, our method generalizes well to in-the-wild lower-body try-on cases despite being trained only on the standard DressCode lower-body setting without ITW fine-tuning. The comparisons in Fig. 9 and Fig. 10 further demonstrate that ToWo-VTON produces more stable body alignment, better identity preservation, and more faithful garment fitting under challenging lower-body and dress try-on scenarios. Additional upper-body results on VITON-HD and DressCode are provided in Fig. 11 and Fig. 12, where our method maintains consistent garment–body alignment while preserving fine-grained garment details. Overall, these qualitative examples show that ToWo-VTON achieves robust try-on synthesis across diverse garment categories, poses, and occlusion patterns.

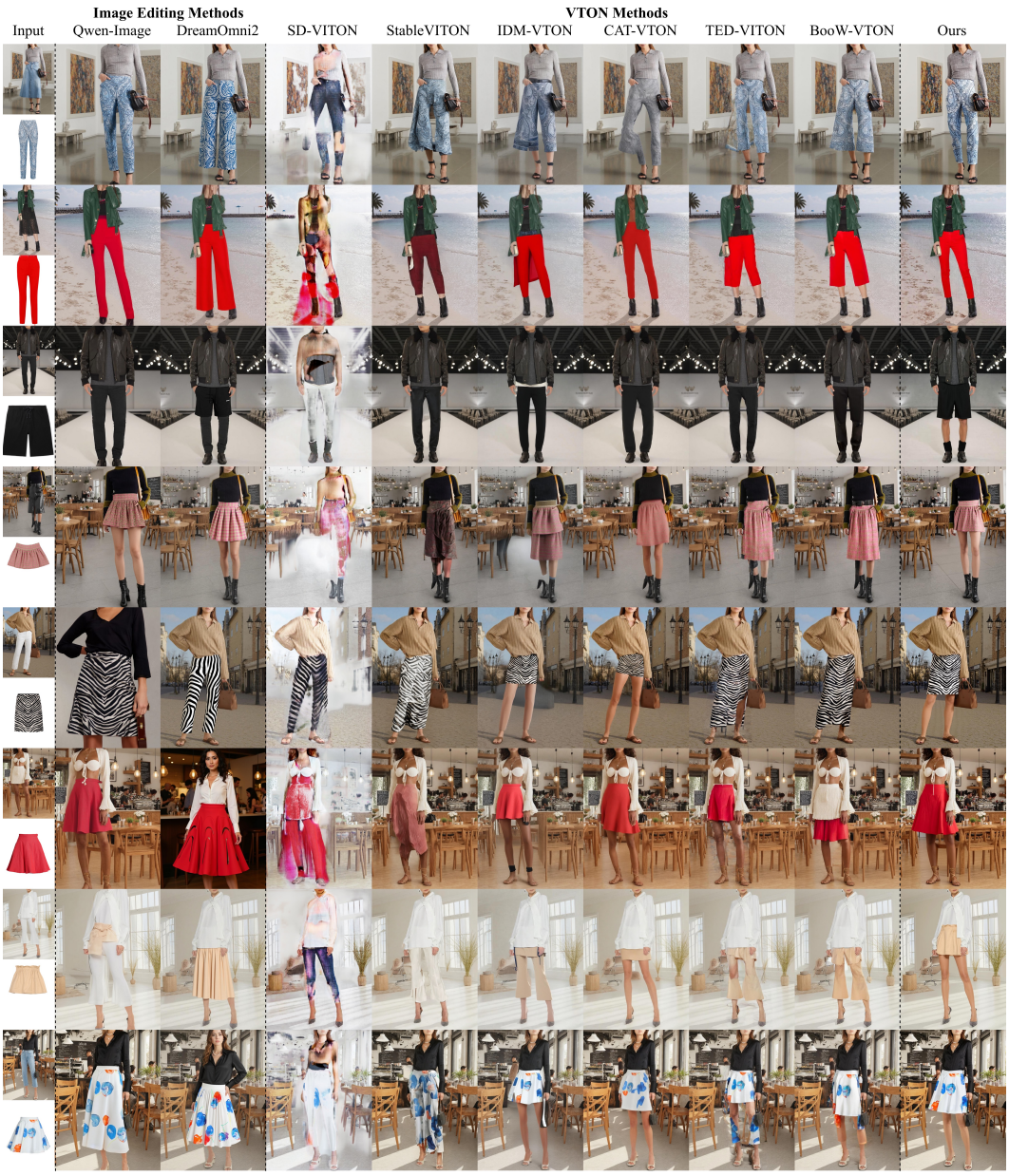

Figure 8: In-the-Wild Lower-Body Try-On Comparison. Despite being trained on the standard DressCode lower-body without ITW fine-tuning, our method demonstrates superior robustness to background variation and better preservation of garment and identity details.



Figure 9: Additional Qualitative Evaluation on Dress-Code Lower-Body dataset. ToWo-VTON achieves significantly more consistent body alignment, human identity retention, and superior preservation of garment spatial details compared to existing approaches.



Figure 10: Additional Qualitative Evaluation on Dress-Code Dresses dataset. ToWo-VTON achieves significantly more consistent body alignment, human identity retention, and superior preservation of garment spatial details compared to existing approaches.



Figure 11: Qualitative comparison on VITON-HD dataset. ToWo-VTON achieves more consistent garment–body alignment and clearer garment details than prior methods.



Figure 12: Qualitative comparison on Dress-Code Upper-Body dataset. ToWo-VTON achieves more consistent garment–body alignment and clearer garment details than prior methods.



Figure 13: ToWo-VTON DressCode Lower-Body Try-on Results. Visual demonstration of garment transfer robustness across diverse poses and clothing types.



Figure 14: ToWo-VTON DressCode Full-Body (Dresses) Try-on Results. Visual demonstration of garment transfer robustness across diverse poses and clothing types.



Figure 15: ToWo-VTON DressCode Upper-Body Try-on Results. Visual demonstration of garment transfer robustness across diverse poses and clothing types.

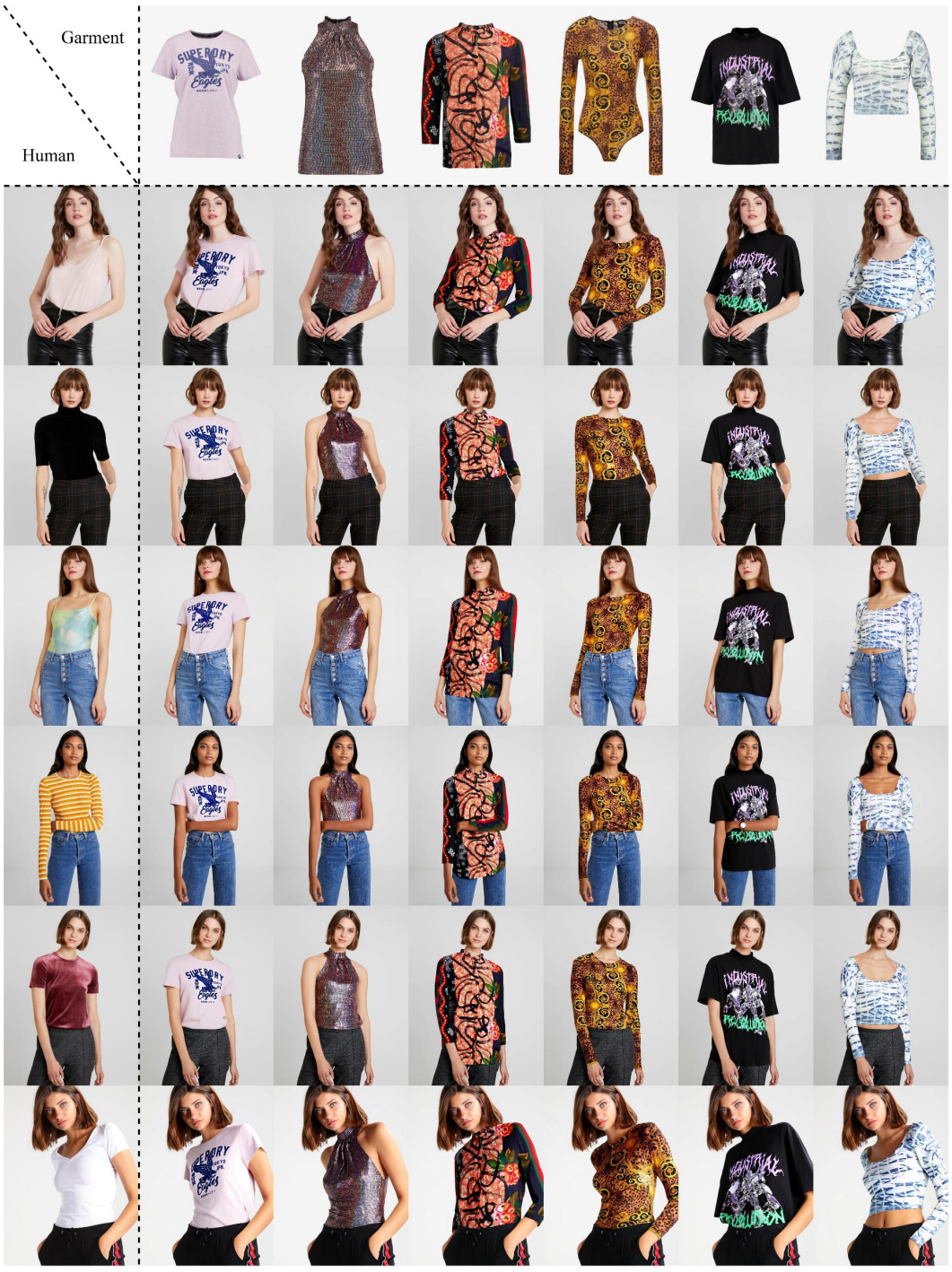

Figure 16: ToWo-VTON DressCode VITON-HD Try-on Results. Visual demonstration of garment transfer robustness across diverse poses and clothing types.