Reference-Free Image Quality Assessment for Virtual Try-On via Human Feedback

Yuki Hirakawa[1,3], Takashi Wada[1], Ryotaro Shimizu[1], Takuya Furusawa[1], Yuki Saito[1], Ryosuke Araki[2], Tianwei Chen[1], Fan Mo[1], and Yoshimitsu Aoki[3]

[1] ZOZO Research
[2] ZOZO Inc.
[3] Keio University

Abstract. As virtual try-on (VTON) systems become increasingly important in fashion e-commerce, there is a growing need for reliable reference-free evaluation methods, since ground-truth images of the same person wearing the target garment are typically unavailable in real-world scenarios. To address this challenge, we propose VTON-IQA, a reference-free framework for human-aligned image quality assessment without requiring ground-truth images. To model human perceptual judgments, we construct VTON-QBench, a large-scale human-annotated benchmark comprising 62,688 try-on images generated by 14 representative VTON models and 431,800 quality annotations collected from 13,838 qualified annotators. To the best of our knowledge, this is the largest dataset to date for human subjective evaluation in VTON. Extensive experiments show that VTON-IQA achieves reliable human-aligned image quality assessment. Moreover, we conduct a comprehensive benchmark evaluation of 14 representative VTON models using VTON-IQA. The dataset and source code are released at `https://github.com/litelightlite/VTON-IQA`.

Keywords: Virtual Try-On · Image Quality Assessment

1 Introduction

Given a person image and a garment image, image-based virtual try-on (VTON) synthesizes a try-on image of the person wearing the target garment. In the fashion e-commerce domain, it has attracted considerable attention as a means of providing an online try-on experience and bridging the gap between online and in-store shopping. Despite its growing practical importance, reliable evaluation remains a fundamental challenge. In real-world deployment, ground-truth images of the same person wearing the target garment are typically unavailable, rendering reference-based metrics such as Structural Similarity Index (SSIM) [33] and Learned Perceptual Image Patch Similarity (LPIPS) [40] impractical. Although large-scale user studies can offer reliable assessments, they are costly, time-consuming, and difficult to deploy in production settings. Distribution-level metrics such as Fréchet Inception Distance (FID) [13] and Kernel Inception Distance (KID) [1], on the other hand, capture dataset-level statistics but fail to

assess the perceptual quality of individual generated images. Recent efforts have introduced virtual try-on evaluation methods that do not rely on ground-truth images [15, 31, 34]. While these methods mark important progress toward task-specific, reference-free evaluation, their alignment with large-scale human judgments remains insufficiently validated. Moreover, the lack of publicly available implementations and standardized benchmarks hinders reproducible evaluation and limits sustained progress in the virtual try-on community.

To address these challenges, we propose VTON-IQA, a reference-free evaluation framework that performs image quality assessment (IQA) of virtual try-on images without relying on ground truth. Given a person image, a garment image, and the corresponding generated try-on image, VTON-IQA predicts a continuous quality score aligned with human subjective evaluations. To model human perceptual judgments, we construct VTON-QBench, a large-scale human-annotated benchmark tailored to virtual try-on images. VTON-QBench consists of 62,688 try-on images generated by 14 representative VTON models and annotated through crowdsourced subjective evaluations. Since crowdsourcing inevitably introduces variability in annotation quality, including inattentive or adversarial responses, we design a dedicated data curation pipeline to quantitatively assess annotator consistency and reliability. This process allows us to identify and exclude unreliable annotators from the final dataset. As a result, VTON-QBench contains 431,800 high-quality annotations collected from 13,838 qualified annotators. To the best of our knowledge, this is the largest dataset to date for human subjective quality evaluation in the virtual try-on domain. With VTON-IQA and VTON-QBench, we provide a scalable and reproducible alternative to large-scale user studies for evaluating virtual try-on quality.

In virtual try-on, quality assessment must account for not only the perceptual quality of the generated image, but also garment fidelity and the preservation of non-target visual elements, such as the person's identity, body shape, pose, and background. This requirement fundamentally distinguishes try-on image assessment from conventional single-image quality assessment, which predicts quality solely from a single input image [32, 36]. Accordingly, evaluating try-on images requires modeling cross-image interactions among the generated try-on image, the input garment image, and the source person image. To this end, we introduce an Interleaved Cross-Attention (ICA) module, which extends standard transformer blocks by inserting a cross-attention layer between the self-attention and MLP layers in the latter half of the blocks. By enabling structured interactions across the try-on, garment, and person representations, ICA allows the model to jointly assess whether garment attributes are faithfully transferred and whether non-target visual elements remain visually consistent. Extensive experiments demonstrate that ICA effectively captures quality factors specific to virtual try-on and achieves strong alignment with human subjective judgments.

Furthermore, we conduct a comprehensive benchmark evaluation of 14 representative virtual try-on models using VTON-IQA. Beyond these baseline evaluations, the public release of VTON-IQA can provide a fair, reproducible, and reference-free evaluation criterion for future virtual try-on methods, serving as a

reproducible alternative to the ad hoc user studies often conducted in individual research projects. We expect that this will contribute to the standardization of quality assessment and further advance research in virtual try-on.

2 Related Work

Image-based Virtual Try-On. Given a person image and a garment image, image-based virtual try-on synthesizes a try-on image of the person wearing the target garment. Early approaches [4, 12, 20, 29] predominantly followed a two-stage pipeline. A clothing-agnostic representation was first constructed by removing garment-related information from the person image, after which the target garment was geometrically aligned to the body using Thin-Plate Spline transformation and fused via a Generative Adversarial Network (GAN) [8]. Subsequent works improved robustness to complex poses and occlusions through human parsing, pose estimation, and enhanced warping–fusion modules, leading to better texture preservation and more natural garment boundaries [4,7,11,20,29,38,41]. More recently, diffusion-based approaches have gained popularity due to their strong ability to reproduce high-frequency details and generate high-resolution images [14, 28]. Early diffusion-based VTON models leveraged pretrained latent diffusion models and incorporated garment features as conditioning signals within U-Net architectures [22, 42]. Later works explored conditioning strategies and architectural refinements. For example, IDM-VTON [5] enhances garment fidelity by disentangling semantic and texture features within attention modules, whereas CatVTON [6] adopts a simplified end-to-end design by spatially concatenating person and garment images. Recent efforts have further transitioned to Diffusion Transformers (DiT) [27], which offer stronger global modeling through self-attention [10,17], enabling improved long-range dependency modeling of garment structure and texture. In parallel, advances in proprietary image editing models [9, 26] demonstrate that high-quality virtual try-on can be achieved in a zero-shot manner using a person image, a garment image, and appropriate textual prompts.

Quality Assessment for Virtual Try-On. Recent works have explored evaluation frameworks for virtual try-on that do not require ground-truth images. VTON-VLLM [31] trains a multimodal large language model using human annotations on LLM-generated critiques of synthesized try-on images, enabling preference-aware evaluation and feedback. However, its focus is on validating and generating textual critiques rather than learning a direct quantitative image-level quality predictor. VTBench [15] introduces a hierarchical benchmark that evaluates virtual try-on results across multiple quality dimensions. Although it incorporates human-aligned signals through LLM-based general aesthetic predictor, it does not use a quality assessment model specifically tailored to virtual try-on. VTONQA [34] is most closely related to our work, as it constructs a human-annotated dataset to train a virtual try-on quality evaluator. However, its relatively limited dataset scale may restrict the robustness of the learned model. In contrast, we construct a large-scale human-annotated benchmark and

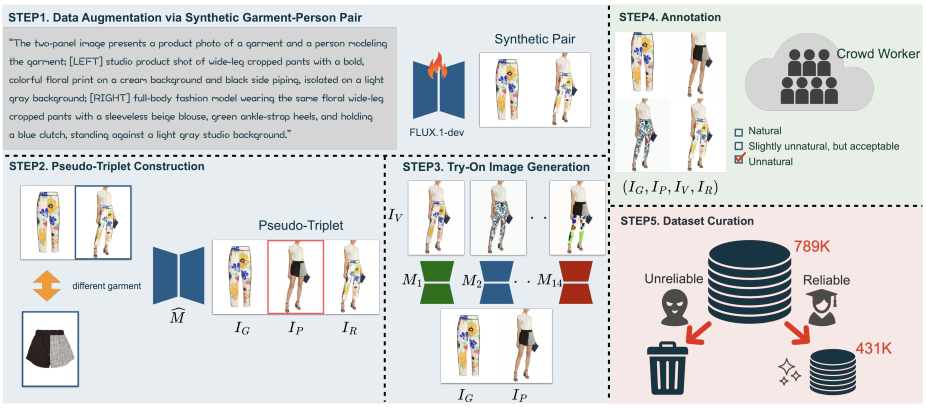

Fig. 1: Overview of the VTON-QBench construction pipeline. VTON-QBench is built through five stages: (1) synthetic garment–person pair augmentation via FLUX.1-dev, (2) pseudo-triplet construction, (3) virtual try-on image generation using 14 representative VTON models, (4) crowdsourced human annotation with reference images, and (5) dataset curation to remove unreliable annotations. This pipeline ensures fashion diversity, controlled evaluation settings, and reliable human-aligned quality labels.

Table 1: Comparison of dataset statistics. VTON-QBench significantly expands the scale of garment–person pairs, try-on images, annotators, and quality annotations compared to prior work, and is publicly available (OSS).

Method	# Pair	# Try-On	# VTON Model	# Annotator	# Annotation	OSS
VTONQA [34]	748	8,132	11	40	N/R	No
VTON-QBench	13,153	62,688	14	13,838	431,800	Yes

learn a reference-free, image quality assessment model aligned with human perceptual judgments.

3 VTON-QBench

To enable human-aligned quality assessment for VTON, we construct VTON-QBench, a large-scale dataset comprising 62,688 try-on images generated by 14 representative models and 431,800 quality annotations collected from 13,838 crowd workers who met reliability criteria. Fig. 1 shows an overview of the VTON-QBench construction pipeline, which consists of five stages: (1) synthetic garment–person pair augmentation via FLUX.1-dev, (2) pseudo-triplet construction, (3) virtual try-on image generation using 14 representative VTON models, (4) crowdsourced human annotation with reference images, and (5) dataset curation to remove unreliable annotations. As shown in Table 1, VTON-QBench is larger in scale than the concurrent work [34] and, to the best of our knowledge, the only publicly available dataset for human subjective evaluation in VTON.

3.1 Image Preparation

Data Augmentation via Synthetic Garment–Person Pairs. VTON-QBench is constructed from the test splits of VITON-HD [4] and Dress Code [23], both of which provide high-quality garment–person image pairs (I_G, I_P). However, due to the limited diversity of their test sets, the constructed pairs lack variation in garment styles, body shapes, poses, and person appearances. To improve coverage beyond the original benchmarks, we augment the data with synthetic garment–person pairs, as illustrated in Fig. 2a. Specifically, we generate additional pairs according to the fashion style taxonomy of Mystylebox [24], covering 24 diverse styles. For this generation process, we employ FLUX.1-dev [2] and train a LoRA specialized for garment–person pair synthesis [16] using an internally collected dataset. This allows us to obtain paired garment and person images with broader variations in appearance while maintaining compatibility between the garment and the target person. Since synthetic generation may introduce inconsistencies between the garment image and the corresponding person image, such as mismatches in category, color, silhouette, or texture, we apply a two-stage verification process. First, we perform automatic attribute-consistency filtering using GPT-4.1 [25] to remove pairs with obvious semantic or visual mismatches. Second, experienced reviewers familiar with fashion images, including researchers in related computer vision fields, manually inspect the remaining pairs to ensure that garment attributes are preserved and that the generated person images are visually plausible. We retain only pairs that pass both stages. As a result, the number of garment–person pairs increases from 6,981 to 13,153, corresponding to approximately 1.9× the original scale. Additional details on the generation procedure are provided in the supplementary material.

Pseudo-Triplet Construction. To enable comparison with existing reference-based metrics such as SSIM [33] and LPIPS [40], we construct pseudo-triplets from existing garment–person pairs $(I_{G'}, I_{P'})$, where $I_{P'}$ is the ground-truth person image wearing garment $I_{G'}$. Given a pair $(I_{G'}, I_{P'})$, we sample a different garment image $I_{G''}$ such that $G'' \neq G'$, and apply a strong virtual try-on model $\widehat{M}$ to generate an intermediate try-on image $I_{V''} = \widehat{M}(I_{G''}, I_{P'})$, in which the original garment G' is replaced with G''. We then define the pseudo-triplet as $(I_G, I_P, I_R) = (I_{G'}, I_{V''}, I_{P'})$, where $I_{G'}$ serves as the target garment image, $I_{V''}$ as the input person image, and $I_{P'}$ as the ground-truth reference image. Given this pseudo-triplet, a target VTON model M generates try-on image $I_V = M(I_{G'}, I_{V''})$, and reference-based metrics can be computed by comparing I_V and $I_{P'}$. This construction converts an otherwise unpaired setting into a pseudo-reference setting, allowing systematic comparison with reference-based metrics. In our experiments, we use Nano Banana Pro [9] as the strong virtual try-on model $\widehat{M}$. Since generating $I_{V''}$ may introduce unintended changes, such as pose shifts, background alterations, or modifications outside the target garment region, we manually filter the generated samples and retain only those in which garment replacement is the primary difference while the person structure and background are preserved.

(a) Synthetic garment–person pairs. **(b)** Distribution of Krippendorff's α.

Fig. 2: Left: Representative synthetic garment–person pairs introduced to enhance fashion diversity. Right: Distribution of Krippendorff's α before and after data curation, showing improved inter-annotator agreement after filtering unreliable annotations.

Try-On Image Generation. We generate virtual try-on images I_V using 14 representative models: VITON-HD [4], HR-VITON [20], LADI-VTON [22], SD-VITON [29], CAT-DM [39], OOTDiffusion [37], IDM-VTON [5], CatVTON [6], CatVTON-FLUX [6], FitDit [17], Any2AnyTryon [10], Qwen-Image-Edit [35], Nano Banana Pro [9], and GPT-Image-1.5 [26]. For each model, we use publicly released pretrained weights and follow the recommended inference settings whenever available.

3.2 Annotation Protocol

For each clothing image I_G, person image I_P, and generated try-on image I_V, we collect independent quality annotations $\mathcal{A}$ from multiple crowd workers. Our preliminary experiments showed that garment length consistency is difficult to assess from the clothing image I_G alone. Therefore, we additionally provide a reference image I_R, where the target garment is worn by a real person, as auxiliary context. Each questionnaire contains 50 evaluation tasks with a fixed structure and rating format. By varying the presented image tuples, we construct 2,139 questionnaires, each of which is assigned to multiple crowd workers to obtain repeated independent evaluations for every sample. Each annotation $a \in \mathcal{A}$ is given on a three-level ordinal scale: (1) unnatural, (2) slightly unnatural but acceptable, and (3) natural. The final subjective quality score $S(I_G, I_P, I_V)$ is computed as the average numerical rating across annotators:

$$S(I_G, I_P, I_V) = \frac{1}{|\mathcal{A}(I_G, I_P, I_V)|} \sum_{a \in \mathcal{A}(I_G, I_P, I_V)} L(a), \tag{1}$$

where $L(a)$ is defined as follows:

$$L(a) = \begin{cases} 1 & \text{if } a = \text{“Unnatural”}, \\ 2 & \text{if } a = \text{“Slightly unnatural, but acceptable”}, \\ 3 & \text{if } a = \text{“Natural”}. \end{cases} \tag{2}$$

3.3 Dataset Curation

Crowdsourced annotation inevitably introduces variability in response quality, including careless or inconsistent responses. To collect reliable annotations, we design a two-stage curation pipeline that removes unreliable annotators and low-agreement annotations. First, we perform annotator-level filtering using five shared dummy tasks with unambiguous answers, which are included among the 50 tasks assigned to each annotator. Annotators who fail these sanity checks are removed. We further exclude annotators who select identical responses for more than 80% of the tasks or disagree with the majority vote in more than 60% of cases. These annotator-level filters reduce the total number of annotations from 1,149,144 to 480,412. Next, we perform questionnaire-level filtering based on inter-rater agreement. Specifically, we compute Krippendorff's α [18] for the three-level ordinal labels within each questionnaire and discard questionnaires with $\alpha \leq 0.4$. This yields 431,800 retained annotations after curation. As shown in Fig. 2b, the average agreement increases from 0.286 to 0.550, and 94.5% of the retained questionnaires achieve $\alpha > 0.4$. Given the inherent variability of subjective evaluation, an α above 0.4 is generally considered acceptable, indicating substantially improved annotation consistency [19].

4 Image Quality Assessment for VTON (VTON-IQA)

Given a garment image I_G, a person image I_P, and a generated try-on image I_V, VTON-IQA predicts a continuous quality score $\hat{s} \in [-1, 1]$ aligned with human perceptual judgments. To effectively assess try-on quality from these inputs, we introduce the Interleaved Cross-Attention (ICA) module, which captures cross-image relationships among the garment, person, and try-on images. This section presents the overall architecture of VTON-IQA, followed by a detailed description of the proposed ICA module.

4.1 Model Architecture

We adopt a three-branch transformer architecture to process the garment image I_G, the person image I_P, and the generated try-on image I_V, as illustrated in Fig. 3. Each image is first embedded into a sequence of patch tokens augmented with a `[CLS]` token:

$$X_m^{(0)} = \phi(I_m) \in \mathbb{R}^{N \times d}, \quad m \in \{G, P, V\}, \tag{3}$$

where N is the number of tokens and d is the embedding dimension. The feature extraction module consists of L transformer blocks. For the first half of the network, standard transformer blocks are applied independently to each branch:

$$\begin{aligned} \widetilde{X}_m^{(\ell)} &= X_m^{(\ell-1)} + \mathrm{SA}^{(\ell)}(X_m^{(\ell-1)}), \\ X_m^{(\ell)} &= \widetilde{X}_m^{(\ell)} + \mathrm{MLP}^{(\ell)}(\widetilde{X}_m^{(\ell)}), \end{aligned} \quad \ell = 1, \ldots, L/2, \quad m \in \{G, P, V\}. \tag{4}$$

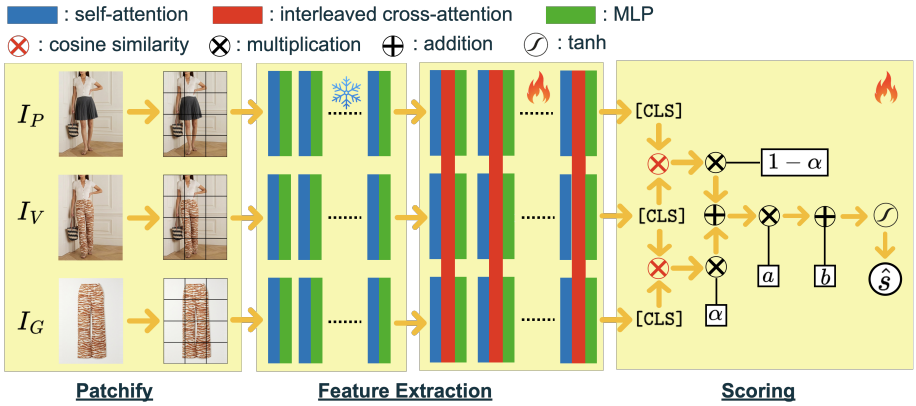

Fig. 3: Architecture of VTON-IQA. The network processes I_G, I_P, and I_V through a three-branch transformer backbone. The first half of layers perform independent feature extraction, while the latter half incorporates Interleaved Cross-Attention (ICA) to explicitly model cross-image interactions. The scoring module aggregates `[CLS]` representations to predict a human-aligned image-level quality score.

In the latter $L/2$ layers, we extend the standard transformer block by introducing the proposed ICA module. Unlike conventional multi-branch designs that symmetrically model all pairwise interactions, ICA is motivated by the observation that virtual try-on quality primarily depends on the consistency between the generated try-on image and its corresponding garment and person images. Specifically, after the self-attention operation, cross-attention is applied between the try-on representation and each of the garment and person representations. In other words, interactions are explicitly modeled for the modality pairs (V, G) and (V, P) in both directions. The cross-attention features are computed as

$$\begin{aligned} C^{(\ell)}_{m\leftarrow n} = \mathrm{CA}^{(\ell)}(Q = \widetilde{X}^{(\ell)}_m, K = \widetilde{X}^{(\ell)}_n, V = \widetilde{X}^{(\ell)}_n), \\ (m, n) \in \{(V, G), (G, V), (V, P), (P, V)\}. \end{aligned} \tag{5}$$

To reflect the asymmetric dependency structure, the try-on representation X_V aggregates contributions from both garment and person features:

$$\widehat{X}^{(\ell)}_V = \widetilde{X}^{(\ell)}_V + C^{(\ell)}_{V\leftarrow G} + C^{(\ell)}_{V\leftarrow P}. \tag{6}$$

In contrast, the garment and person branches only incorporate information from the try-on branch:

$$\widehat{X}^{(\ell)}_G = \widetilde{X}^{(\ell)}_G + C^{(\ell)}_{G\leftarrow V}, \quad \widehat{X}^{(\ell)}_P = \widetilde{X}^{(\ell)}_P + C^{(\ell)}_{P\leftarrow V}. \tag{7}$$

This asymmetric interaction design emphasizes that the quality judgment is fundamentally centered on the generated try-on image, which must be evaluated with respect to both garment fidelity and preservation of non-target visual

elements. By explicitly modeling $V \leftrightarrow G$ and $V \leftrightarrow P$ interactions while avoiding unnecessary $G \leftrightarrow P$ coupling, ICA provides structured relational modeling tailored to virtual try-on quality assessment. After the final layer L, we extract the `[CLS]` token from each branch to obtain compact global representations c_G, c_P, and c_V for the garment, person, and try-on images, respectively. We first compute an intermediate relational score $\tilde{s}$ as a weighted combination of cosine similarities between the try-on representation and the garment/person representations:

$$\tilde{s} = \alpha \frac{c_G^\top c_V}{\|c_G\| \, \|c_V\|} + (1-\alpha) \frac{c_P^\top c_V}{\|c_P\| \, \|c_V\|}, \tag{8}$$

where $\alpha \in [0,1]$ is a learnable scalar parameter that adaptively balances the relative importance of garment consistency and preservation of non-target regions. This formulation explicitly captures two complementary aspects of virtual try-on quality: (i) garment consistency via the similarity between c_G and c_V, and (ii) preservation of non-target visual elements via the similarity between c_P and c_V. By learning α, the model dynamically weights these two relational components, allowing the overall quality score to reflect their relative significance in human perceptual judgment. Finally, the predicted score is obtained through a learnable affine transformation followed by a tanh activation:

$$\hat{s} = \tanh(a\tilde{s} + b), \tag{9}$$

where a and b are learnable scalar parameters. The tanh function constrains the score to the bounded interval $[-1,1]$, which improves numerical stability, enhances interpretability, and enables consistent comparison across models.

4.2 Loss Function

Human ratings for virtual try-on images can exhibit high variance on an absolute scale, whereas relative preferences between two outputs for the same person–garment pair are often more consistent. We therefore optimize a joint objective that combines pairwise preference learning with score regression. Inspired by the Bradley–Terry model [3], we model pairwise preferences between two try-on results, I_{V_i} and I_{V_j}, generated from the same person–garment pair (I_G, I_P). The predicted preference probability is defined as

$$p_\theta := P_\theta(I_{V_i} \succ I_{V_j} \mid I_G, I_P) = \sigma\left(\frac{\Psi_\theta(I_G, I_P, I_{V_i}) - \Psi_\theta(I_G, I_P, I_{V_j})}{\tau}\right), \tag{10}$$

where $\sigma(\cdot)$ denotes the sigmoid function, $\Psi_\theta(\cdot)$ is the predicted quality score, and τ is a temperature parameter. Similarly, given the human evaluation scores S_i and S_j assigned to I_{V_i} and I_{V_j}, the empirical human preference probability is defined as

$$q_{ij} := P_{\text{human}}(I_{V_i} \succ I_{V_j} \mid I_G, I_P) = \sigma\left(\frac{S_i - S_j}{\tau}\right). \tag{11}$$

The overall objective is given by the soft-label cross-entropy combined with a score regression term:

$$\mathcal{L}_\theta = -\, q_{ij} \log p_\theta - (1 - q_{ij}) \log(1 - p_\theta) + \sum_{k \in \{i,j\}} \|\Psi_\theta(I_G, I_P, I_{V_k}) - S_k\|_2^2 . \quad (12)$$

The first term aligns the predicted pairwise preference distribution with the empirical human preference distribution via soft-label cross-entropy, while the second term enforces consistency between the predicted quality scores and the corresponding human ratings.

5 Experiments

We train VTON-IQA on VTON-QBench and evaluate it on the held-out test set against conventional reference-based metrics and a zero-shot baseline model. We report correlation with human ratings and pairwise ranking accuracy, and conduct ablation studies to quantify the effect of the proposed ICA module. Finally, we apply VTON-IQA to benchmark 14 representative VTON models on Dress Code and VITON-HD under both paired and unpaired settings.

5.1 Experimental Setup

Implementation Details. We build VTON-IQA on top of DINOv3 ViT-L/16 [30]. To integrate the proposed ICA module, we augment the last 12 transformer blocks with ICA layers and fine-tune both the original parameters of these blocks and the inserted ICA layers, while freezing the parameters of the first 12 layers. The model is optimized using AdamW [21] with a batch size of 16 and a learning rate of 1×10^{-4}. We employ early stopping based on the validation loss, selecting the checkpoint with the lowest validation loss if no improvement is observed for three consecutive epochs. Training is conducted on a single NVIDIA A100 (40GB) GPU using bfloat16 mixed-precision.
Dataset. We split VTON-QBench into 43,948 training, 5,702 validation, and 13,038 test samples, with disjoint person and garment identities across splits. The quality scores are normalized to $[-1.0, 1.0]$ using Min–Max normalization followed by a linear transformation, with the normalization parameters estimated from the training set and applied consistently to the validation and test sets.
Baselines. We compare our method with representative reference-based metrics, including SSIM [33] and LPIPS [40]. As a strong reference-free baseline, we include zero-shot DINOv3 [30], in which garment, person, and try-on images are processed independently and the final score is computed using the cosine similarity formulation in Eq. (8), with $\alpha = 0.5$. To assess the effectiveness of ICA, we further evaluate VTON-IQA with and without ICA.
Metrics. We evaluate prediction performance by measuring agreement with human subjective scores. We report Pearson's Linear Correlation Coefficient (PLCC, ρ_{PLCC}) and Spearman's Rank Correlation Coefficient (SRCC, ρ_{SRCC})

Table 2: Comparison with baseline performance. Note that LPIPS is originally a "lower-is-better" metric; for consistency, we compute all metrics on its negated values. For SSIM, LPIPS, and zero-shot DINOv3, ρ_{PLCC} and R^2 are not reported, as their score scales are not directly comparable to human subjective ratings.

Scorer	$\rho_{\text{SRCC}} \uparrow$	$\rho_{\text{PLCC}} \uparrow$	$R^2 \uparrow$	$A_{\text{macro}} \uparrow$	$A_{\text{micro}} \uparrow$	reference-free	fine-tuned
SSIM [33]	0.150	–	–	0.596	0.593	×	×
LPIPS [40]	0.406	–	–	0.701	0.695	×	×
DINOv3 [30]	0.244	–	–	0.637	0.641	✓	×
VTON-IQA w/o ICA	0.617	0.615	0.372	0.722	0.747	✓	✓
VTON-IQA	**0.750**	**0.751**	**0.553**	**0.781**	**0.790**	✓	✓

Table 3: Comparison with human performance. We estimate human performance by randomly splitting test-set annotators into two groups and treating their aggregated scores as ground truth and predictions, respectively. For fair comparison, VTON-IQA is evaluated against labels aggregated from half of the test-set annotators.

Scorer	$\rho_{\text{SRCC}} \uparrow$	$\rho_{\text{PLCC}} \uparrow$	$R^2 \uparrow$	$A_{\text{macro}} \uparrow$	$A_{\text{micro}} \uparrow$
Ours	0.699 ± 0.002	0.700 ± 0.002	0.489 ± 0.004	0.771 ± 0.003	0.783 ± 0.002
Human	$\mathbf{0.760 \pm 0.004}$	$\mathbf{0.762 \pm 0.004}$	$\mathbf{0.536 \pm 0.008}$	$\mathbf{0.782 \pm 0.002}$	$\mathbf{0.791 \pm 0.002}$

over all generated try-on images to assess global score consistency with human ratings. PLCC measures linear agreement, while SRCC evaluates whether the predicted scores preserve the ranking induced by human judgments. We also report the coefficient of determination (R^2) to assess regression accuracy with respect to absolute human scores. We further evaluate within-pair ranking consistency using pairwise accuracy at the garment–person pair level. For each garment–person pair M, all generated try-on images are compared pairwise, resulting in $\binom{N_M}{2}$ comparisons, where N_M is the number of images for pair M. A comparison is counted as correct if the predicted scores preserve the relative ordering of the corresponding human scores. We report micro-averaged pairwise accuracy (A_{micro}), computed over all pairwise comparisons, and macro-averaged pairwise accuracy (A_{macro}), computed per pair and then averaged. Because N_M is relatively small and varies across pairs, A_{macro} can be unstable; thus, we use A_{micro} as the main pairwise metric and A_{macro} as a complementary metric.

5.2 Quantitative Results

Comparison with Baselines. Table 2 summarizes the evaluation results on the test set of VTON-QBench. For SSIM, LPIPS, and zero-shot DINOv3, ρ_{PLCC} and R^2 are not reported, as their score scales are not directly comparable to human subjective ratings, rendering these statistics less meaningful. Note that LPIPS is originally a "lower-is-better" metric; for consistency, we compute all correlation and regression metrics on its negated values so that all metrics follow a "higher-

Fig. 4: Qualitative results. From left to right: garment image, target person image, generated try-on results (columns 3–7), and ground-truth image. The top-right value shows the human score, and the top-left black box indicates each metric's ranking.

is-better" convention. Among reference-based metrics, SSIM exhibits weak alignment with human judgment, achieving low SRCC and pairwise accuracy. LPIPS performs better than SSIM but still shows a substantial gap compared to ours. The comparison between zero-shot DINOv3 and VTON-IQA without ICA highlights the importance of task-specific training on VTON-QBench, while further performance gains obtained by incorporating garment–person interaction modeling demonstrate the effectiveness of the proposed ICA module. The full VTON-IQA achieves the highest scores across all reported metrics, with $\rho_{\mathrm{SRCC}} = 0.750$, $\rho_{\mathrm{PLCC}} = 0.751$, and the best macro and micro pairwise accuracy.

Comparison with Human. To estimate human performance, we split the annotators in the test set into two disjoint groups and compute independent quality scores from each group. One group is treated as ground-truth labels, while the other serves as predictions. This random partitioning is repeated 10 times, and we report the mean and standard deviation of the metrics across runs. For fair comparison with human performance, VTON-IQA is evaluated against the same ground-truth group in each run, while being trained on quality scores aggregated from all available annotators in the training set. Table 3 presents the comparison between human performance and our model. For correlation-based metrics (ρ_{SRCC} and ρ_{PLCC}) and the coefficient of determination (R^2), a noticeable gap remains between our model and human performance, indicating room for further improvement in capturing fine-grained perceptual alignment. In contrast, our method achieves performance close to the human level in terms of macro accuracy (A_{macro}) and micro accuracy (A_{micro}). This suggests that, for

Table 4: Evaluation results of VTON methods on Dress Code [23]. Since several GAN-based models support only upper-body garments, their evaluation on Dress Code is restricted to upper-body categories and reported in gray for reference. **Bold** and underlined values indicate the best and second-best results for each metric, respectively.

	paired					*unpaired*		
Method	Ours ↑	FID ↓	KID ↓	SSIM ↑	LPIPS ↓	Ours ↑	FID ↓	KID ↓
Proprietary image edit model								
Nano Banana Pro [9]	**0.305**	**1.837**	0.471	0.917	**0.045**	**0.295**	5.179	1.048
GPT-Image-1.5 [26]	0.237	4.270	0.729	0.824	0.170	0.205	5.621	0.845
DiT-based diffusion								
Qwen-Image-Edit [35]	-0.094	6.827	4.541	**0.928**	0.066	-0.081	7.495	3.477
CatVTON-FLUX [6]	0.014	3.093	**0.291**	0.890	0.122	-0.132	4.914	0.767
FitDit [17]	0.219	3.374	0.623	0.899	0.083	0.153	**4.771**	**0.642**
Any2AnyTryon [10]	-0.045	4.312	1.347	0.895	0.133	-0.236	6.359	1.800
U-Net-based diffusion								
LADI-VTON [22]	-0.562	6.608	2.907	0.893	0.150	-0.694	9.131	4.414
CAT-DM [39]	-0.412	5.643	2.229	0.881	0.143	-0.741	8.189	3.408
OOTDiffusion [37]	-0.230	4.571	1.044	0.888	0.080	-0.511	8.593	2.981
IDM-VTON [5]	0.141	3.223	0.797	0.911	0.058	-0.349	4.931	1.095
CatVTON [6]	-0.081	5.306	1.841	0.882	0.110	-0.232	6.927	2.015
GAN-based								
VITON-HD [4]	-0.933	94.605	90.826	0.854	0.224	-0.933	96.261	94.204
HR-VITON [20]	-0.835	20.873	8.767	0.917	0.109	-0.841	22.626	9.337
SD-VITON [29]	-0.868	17.494	5.736	0.911	0.107	-0.874	19.575	6.798

pairwise quality comparisons of individual try-on images, the proposed model demonstrates human-comparable decision consistency.

5.3 Qualitative Results

Fig. 4 presents representative qualitative comparisons. From left to right, the columns show the garment image, the target person image, the generated try-on results (columns 3–7), and the ground-truth image. The try-on results are arranged in ascending order of human subjective scores. For each try-on image, the value in the top-right corner denotes the human score, while the black box in the top-left corner indicates the rank of the five try-on images assigned by each metric. In the first row, only the top human-rated sample faithfully preserves both garment length and printed text details. Both our method and LPIPS correctly rank this sample as the top result, demonstrating sensitivity to fine-grained garment consistency. The second and third rows highlight cases involving structural changes. The second row contains pose variations, while the third row exhibits differences in zoom level, which often arise in mask-free try-on models without

Table 5: Evaluation results of VTON methods on VITON-HD [4]. **Bold** and underlined values indicate the best and second-best results for each metric, respectively.

	paired					*unpaired*		
Method	Ours ↑	FID ↓	KID ↓	SSIM ↑	LPIPS ↓	Ours ↑	FID ↓	KID ↓
Proprietary image edit model								
Nano Banana Pro [9]	**0.303**	**3.318**	0.707	**0.904**	**0.052**	**0.315**	10.309	2.454
GPT-Image-1.5 [26]	0.255	9.613	3.500	0.768	0.220	0.234	12.801	4.631
DiT-based diffusion								
Qwen-Image-Edit [35]	0.137	6.195	2.849	0.882	0.071	0.087	10.706	3.030
CatVTON-FLUX [6]	0.207	5.470	0.413	0.872	0.132	-0.025	9.084	0.878
FitDit [17]	0.230	8.048	1.232	0.843	0.144	0.189	9.893	1.355
Any2AnyTryon [10]	0.074	9.043	3.271	0.860	0.169	-0.082	12.018	3.892
U-Net-based diffusion								
LADI-VTON [22]	-0.756	15.404	7.416	0.847	0.202	-0.864	21.515	12.333
CAT-DM [39]	-0.419	9.154	2.340	0.842	0.171	-0.511	11.774	3.141
OOTDiffusion [37]	0.018	5.915	**0.290**	0.845	0.114	-0.142	**9.064**	**0.543**
IDM-VTON [5]	0.151	6.007	0.775	0.865	0.101	0.039	9.093	1.119
CatVTON [6]	-0.163	9.082	2.710	0.833	0.196	-0.266	11.199	2.824
GAN-based								
VITON-HD [4]	-0.512	9.450	1.956	0.877	0.130	-0.559	11.535	2.437
HR-VITON [20]	-0.420	9.379	2.578	0.883	0.118	-0.504	11.809	2.921
SD-VITON [29]	-0.368	7.746	1.567	0.886	0.112	-0.479	10.412	1.912

strict pose or scale constraints. Since SSIM and LPIPS rely on pixel- or feature-level alignment with the reference image, they tend to over-penalize such global transformations and produce rankings that deviate from human judgments. In contrast, VTON-IQA remains consistent with human perception, demonstrating robustness to pose and scale variations while still capturing try-on-specific quality differences.

5.4 Quality Assessment Results for VTON Models

We conduct a comprehensive evaluation of 14 representative VTON models under both *paired* settings (where ground-truth try-on images are available) and *unpaired* settings (where no reference images are provided). In the paired setting, reference-based metrics such as SSIM and LPIPS are applicable, whereas unpaired evaluation typically relies on distribution-level metrics including FID and KID. As a reference-free method, VTON-IQA can be applied consistently in both settings. All experiments are performed on the original test sets of Dress Code and VITON-HD, with results reported in Tables 4–5. Across both datasets, proprietary image editing models achieve the highest VTON-IQA scores. Nano

Banana Pro ranks first in all configurations, followed by GPT-Image-1.5, suggesting stronger alignment with human subjective ratings. In contrast, conventional full-reference metrics yield different rankings. Under SSIM and LPIPS, GPT-Image-1.5 performs substantially worse, particularly on VITON-HD, where it ranks below earlier GAN-based models. As discussed in Section 5.3, these metrics penalize global structural variations such as pose and zoom changes. To further validate this hypothesis, we randomly sample 30 results per model and find that only one Nano Banana Pro sample exhibits noticeable pose or zoom variation, compared to 19 for GPT-Image-1.5, thereby accounting for the lower SSIM/LPIPS scores. More broadly, several diffusion models achieve competitive FID/KID scores yet rank lower than the proprietary models under VTON-IQA, indicating that distribution-level similarity does not necessarily reflect perceptual quality in virtual try-on.

6 Conclusion

We presented VTON-IQA, a reference-free, human-aligned image quality assessment framework for virtual try-on. To support reliable learning and evaluation, we constructed VTON-QBench, a large-scale human-annotated benchmark comprising 62,688 try-on images and 431,800 quality annotations from 13,838 qualified annotators. To ensure the reliability of these annotations, we developed a dedicated curation pipeline that quantitatively assesses annotator consistency and filters unreliable responses and annotators. This process provides reliable subjective supervision for training VTON-IQA to predict human-aligned quality scores from try-on, garment, and person images. By introducing an Interleaved Cross-Attention (ICA) module, VTON-IQA explicitly models interactions among these inputs, leading to stronger alignment with human subjective judgments. With VTON-IQA, we further conduct a comprehensive benchmark evaluation of 14 representative VTON models under both *paired* and *unpaired* settings, revealing that recent proprietary image-editing models achieve higher quality scores than OSS models. Meanwhile, conventional metrics such as SSIM, LPIPS, and FID/KID often fail to reflect these improvements, highlighting the gap between pixel- or distribution-level similarity and human-perceived try-on quality. Beyond our baseline benchmark, the public release of VTON-IQA can provide a fair, reproducible, and reference-free evaluation criterion for future VTON methods, serving as a scalable alternative to the ad hoc user studies often conducted in individual research projects. Overall, VTON-IQA and VTON-QBench support the standardization of quality assessment and sustained progress in the virtual try-on community.

Acknowledgements

We thank Hideya Tanaka, Hiroshige Matsushita, and Motoaki Nakai for their support in building and maintaining the data collection infrastructure.

References

1. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying MMD GANs. In: The Sixth International Conference on Learning Representations (2018)
2. Black Forest Labs: FLUX. `https://github.com/black-forest-labs/flux` (2024)
3. Bradley, R.A., Terry, M.E.: Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. Biometrika **39**(3/4), 324–345 (1952)
4. Choi, S., Park, S., Lee, M., Choo, J.: VITON-HD: High-Resolution Virtual Try-On via Misalignment-Aware Normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14126–14135 (2021)
5. Choi, Y., Kwak, S., Lee, K., Choi, H., Shin, J.: Improving Diffusion Models for Authentic Virtual Try-on in the Wild. In: Computer Vision – ECCV 2024. pp. 206–235 (2025)
6. Chong, Z., Dong, X., Li, H., shiyue Zhang, Zhang, W., Zhao, H., xujie zhang, Jiang, D., Liang, X.: CatVTON: Concatenation Is All You Need for Virtual Try-On with Diffusion Models. In: The Thirteenth International Conference on Learning Representations (2025)
7. Ge, Y., Song, Y., Zhang, R., Ge, C., Liu, W., Luo, P.: Parser-Free Virtual Try-on via Distilling Appearance Flows. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8481–8489 (2021)
8. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Nets. In: Advances in Neural Information Processing Systems. vol. 27, pp. 2672–2680 (2014)
9. Google: Nano Banana Pro (2025), accessed: 2026-02-11
10. Guo, H., Zeng, B., Song, Y., Zhang, W., Liu, J., Zhang, C.: Any2AnyTryon: Leveraging Adaptive Position Embeddings for Versatile Virtual Clothing Tasks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19085–19096 (2025)
11. Han, X., Hu, X., Huang, W., Scott, M.R.: ClothFlow: A Flow-Based Model for Clothed Person Generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10471–10480 (2019)
12. Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S.: VITON: An Image-Based Virtual Try-On Network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7543–7552 (2018)
13. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In: Advances in Neural Information Processing Systems. vol. 30, pp. 6629–6640 (2017)
14. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851 (2020)
15. Hu, X., Yujie, L., Luo, D., Xu, P., Zhang, J., Zhu, J., Wang, C., Fu, Y.: VT-Bench: Comprehensive Benchmark Suite Towards Real-World Virtual Try-on Models. ArXiv **abs/2505.19571** (2025)
16. Huang, L., Wang, W., Wu, Z., Shi, Y., Dou, H., Liang, C., Feng, Y., Liu, Y., Zhou, J.: In-Context LoRA for Diffusion Transformers. ArXiv **abs/2410.23775** (2024)
17. Jiang, B., Hu, X., Luo, D., He, Q., Xu, C., Peng, J., Zhang, J., Wang, C., Wu, Y., Fu, Y.: FitDiT: Advancing the Authentic Garment Details for High-fidelity Virtual Try-on. ArXiv **abs/2411.10499** (2024)
18. Krippendorff, K.: Computing Krippendorff's Alpha-Reliability. Tech. rep., University of Pennsylvania (2011)

19. Ku, M., Li, T., Zhang, K., Lu, Y., Fu, X., Zhuang, W., Chen, W.: ImagenHub: Standardizing the evaluation of conditional image generation models. In: The Twelfth International Conference on Learning Representations (2024)
20. Lee, S., Gu, G., Park, S., Choi, S., Choo, J.: High-Resolution Virtual Try-On with Misalignment and Occlusion-Handled Conditions . In: ECCV 2022: 17th European Conference, 2022, Proceedings. pp. 204–219 (2022)
21. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization. In: The Seventh International Conference on Learning Representations (2019)
22. Morelli, D., Baldrati, A., Cartella, G., Cornia, M., Bertini, M., Cucchiara, R.: LaDI-VTON: Latent Diffusion Textual-Inversion Enhanced Virtual Try-On. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 8580–8589 (2023)
23. Morelli, D., Fincato, M., Cornia, M., Landi, F., Cesari, F., Cucchiara, R.: Dress Code: High-Resolution Multi-category Virtual Try-On. In: Computer Vision – ECCV 2022. pp. 345–362 (2022)
24. Mystylebox: 24 Types of Fashion Styles : Key Elements and Defining Looks. https://www.mystylebox.com/pages/24-types-of-fashion-styles-explained (2026), accessed: 2026-03-01
25. OpenAI: GPT-4 Technical Report. Tech. rep., OpenAI (2023)
26. OpenAI: GPT-Image-1.5 (2025), accessed: 2026-02-11
27. Peebles, W., Xie, S.: Scalable Diffusion Models with Transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4195–4205 (2023)
28. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
29. Shim, S.H., Chung, J., Heo, J.P.: Towards Squeezing-Averse Virtual Try-On via Sequential Deformation. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 4856–4863 (2024)
30. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3. ArXiv **abs/2508.10104** (2025)
31. Wan, S., Chen, J., Cai, Q., Pan, Y., Yao, T., Mei, T.: VTON-VLLM: Aligning Virtual Try-On Models with Human Preferences. In: Advances in Neural Information Processing Systems. vol. 38, pp. 147641–147664 (2025)
32. Wang, J., Chan, K.C., Loy, C.C.: Exploring CLIP for Assessing the Look and Feel of Images. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 2555–2563 (2023)
33. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
34. Wei, X., Wu, S., Xu, Z., Li, Y., Duan, H., Min, X., Zhai, G.: VTONQA: A Multi-Dimensional Quality Assessment Dataset for Virtual Try-on. ArXiv **abs/2601.02945** (2026)
35. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., mei Li, D., Zhang, H., Meng, H., Wei, H., Ni, J.L., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X.X., Wang, Y., Zhang,

Y., Zhu, Y.A., Wu, Y., Cai, Y.J., Liu, Z.Y.: Qwen-Image Technical Report. Tech. rep., Qwen Team (2025)
36. Wu, H., Zhang, Z., Zhang, W., Chen, C., Liao, L., Li, C., Gao, Y., Wang, A., Zhang, E., Sun, W., Yan, Q., Min, X., Zhai, G., Lin, W.: Q-ALIGN: Teaching LMMs for Visual Scoring via Discrete Text-Defined Levels. In: Proceedings of the 41st International Conference on Machine Learning (2024)
37. Xu, Y., Gu, T., Chen, W., Chen, A.: OOTDiffusion: Outfitting Fusion based Latent Diffusion for Controllable Virtual Try-On. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 8996–9004 (2025)
38. Yang, H., Zhang, R., Guo, X., Liu, W., Zuo, W., Luo, P.: Towards Photo-Realistic Virtual Try-On by Adaptively Generating-Preserving Image Content. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
39. Zeng, J., Song, D., Nie, W., Tian, H., Wang, T., Liu, A.A.: CAT-DM: Controllable Accelerated Virtual Try-on with Diffusion Model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8372–8382 (2024)
40. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018)
41. Zhenyu, X., Zaiyu, H., Xin, D., Fuwei, Z., Haoye, D., Xijin, Z., Feida, Z., Xiaodan, L.: 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 (2023)
42. Zhu, L., Yang, D., Zhu, T., Reda, F., Chan, W., Saharia, C., Norouzi, M., Kemelmacher-Shlizerman, I.: TryOnDiffusion: A Tale of Two UNets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4606–4615 (2023)