VTONQA: A Multi-Dimensional Quality Assessment Dataset for Virtual Try-on

Xinyi Wei*
moj-will@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Sijing Wu*†
wusijing@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Zitong Xu
xuzitong@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Yunhao Li
lyhsjtu@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Huiyu Duan
huiyuduan@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Jia Wang
jiawang@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Ning Liu
ningliu@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

arXiv:2601.02945v2 [cs.CV] 19 Aug 2026

Abstract

With the rapid development of e-commerce and digital fashion, image-based virtual try-on (VTON) has attracted increasing attention. However, existing VTON models often suffer from artifacts such as garment distortion and body inconsistency, highlighting the need for reliable quality evaluation of VTON-generated images. To this end, we construct **VTONQA**, the first multi-dimensional quality assessment dataset specifically designed for VTON, which contains 8,132 images generated by 11 representative VTON models, along with 24,396 mean opinion scores (MOSs) across three evaluation dimensions (*i.e.*, clothing fit, body compatibility, and overall quality). Based on VTONQA, we benchmark both VTON models and a diverse set of image quality assessment (IQA) metrics, revealing the limitations of existing methods and highlighting the value of the proposed dataset. We believe that the VTONQA dataset and corresponding benchmarks will provide a solid foundation for perceptually aligned evaluation, benefiting both the development of quality assessment methods and the advancement of VTON models. The dataset we proposed in this paper is publicly available at: https://huggingface.co/datasets/weixiny0408/vtonqa.

CCS Concepts

• **Computing methodologies** → **Computer vision**; **Computer vision tasks**.

Keywords

Virtual try-on, dataset, benchmark, subjective experiment, image quality assessment

*These authors contributed equally to this work.
†Corresponding author.

This work is licensed under a Creative Commons Attribution 4.0 International License.
MM '26, Rio de Janeiro, Brazil
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3838680

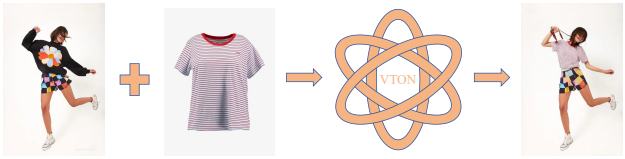


Figure 1: Illustration of the image-based virtual try-on pipeline.

ACM Reference Format:
Xinyi Wei, Sijing Wu, Zitong Xu, Yunhao Li, Huiyu Duan, Jia Wang, and Ning Liu. 2026. VTONQA: A Multi-Dimensional Quality Assessment Dataset for Virtual Try-on. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3767308.3838680

1 Introduction

Image-based virtual try-on (VTON) [6, 18, 46] enables realistic visualization of garments on human images (see Figure 1), with applications in e-commerce, virtual reality, and digital fashion. However, current VTON models often generate low-quality results with artifacts such as blurred faces and garments, distorted body structures, and incorrect clothing transfer, degrading user experience and practical utility. Therefore, effective evaluation of VTON-generated images is essential for assessing perceptual quality, benchmarking VTON models, and guiding model improvement.

Existing evaluations of VTON-generated images mainly rely on objective metrics, including the distribution-based measure Fréchet Inception Distance (FID) [14], the perceptual similarity metric Learned Perceptual Image Patch Similarity (LPIPS) [49], and traditional pixel-level criteria such as Structural Similarity Index (SSIM) [35] and Peak Signal-to-Noise Ratio (PSNR). However, these objective metrics often exhibit weak correlation with human perception, highlighting the importance of subjective quality assessment. In contrast, visual quality assessment (QA) methods [8, 9, 24, 34, 38, 42, 44, 45, 50] typically learn a network to regress quality scores based on human-annotated datasets and

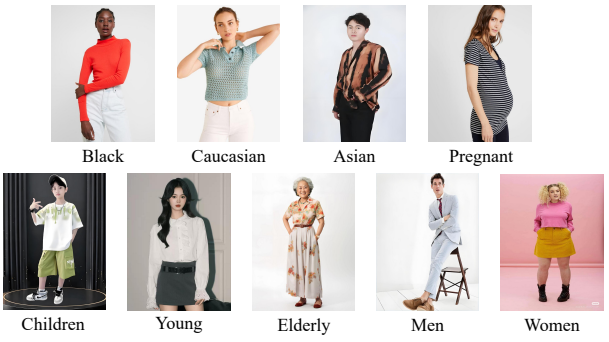


Figure 2: Representative examples of the nine reference person categories in VTONQA, illustrating the diversity of body types and dressing scenarios included in the dataset.



Figure 3: Representative examples of the eight clothing categories included in VTONQA, covering upper-body, lower-body, and full-body garments.

are inherently aligned with human perception. However, existing QA datasets are primarily designed for natural images or specific Artificial Intelligence (AI)-generated images, and none of them specifically target the VTON task. Due to the substantial differences between VTON-generated images and natural images in distortion types, as well as differences in reference information formats and distortion characteristics compared to other AI-generated content [36, 42, 45], QA models trained on these datasets often perform poorly on VTON-generated images, highlighting the need for a VTON-specific QA dataset. However, existing quality assessment methods typically provide a single holistic quality score, which is insufficient to characterize the diverse failure modes of VTON-generated images. In contrast, VTON quality depends on multiple complementary aspects, including garment transfer, body compatibility, and overall perceptual realism. Therefore, a multi-dimensional quality assessment protocol is essential for more fine-grained and interpretable evaluation.

To bridge this gap, we construct **VTONQA**, the first large-scale multi-dimensional QA dataset for VTON-generated images, comprising 8,132 images from 11 representative VTON models and 24,396 mean opinion scores (MOSs) across three dimensions: clothing fit, body compatibility, and overall quality. Specifically, the VTON-generated images are synthesized by applying try-on garments from 8 categories to 184 reference person images spanning 9 categories.

The VTON models include classical warp-based [6, 13, 31, 46], diffusion-based [7, 18, 28, 41, 47], and closed-source methods [19, 26]. Subsequently, 40 subjects annotate images across three evaluation dimensions, with quality ensured through standardized training and supervision by image processing researchers. Based on VTONQA, we benchmark 11 representative VTON models and 17 IQA metrics. All VTON models are evaluated in an inference-only setting without fine-tuning or retraining on VTONQA. For quality assessment, we include full-reference and no-reference IQA methods, covering traditional and deep learning approaches, and show that models trained on VTONQA achieve stronger correlations with human judgments.

We hope that VTONQA, together with its comprehensive benchmarks and subjective annotations, will facilitate the development of objective quality assessment methods for VTON-generated images that better align with human perception and advance robust, realistic VTON models.

In summary, the main contributions are: (1) We present the first comprehensive subjective quality assessment study of VTON-generated images, providing insights into human perceptual preferences across diverse try-on scenarios. (2) We build the first large-scale, multi-dimensional quality assessment dataset for VTON-generated images, containing 8,132 images and 24,396 MOS annotations across three dimensions (*i.e.*, clothing fit, body compatibility, and overall quality), enabling fine-grained evaluation beyond a single score. (3) We benchmark 11 representative VTON models and 17 IQA metrics on VTONQA, revealing the limitations of existing metrics and the importance of subjective supervision.

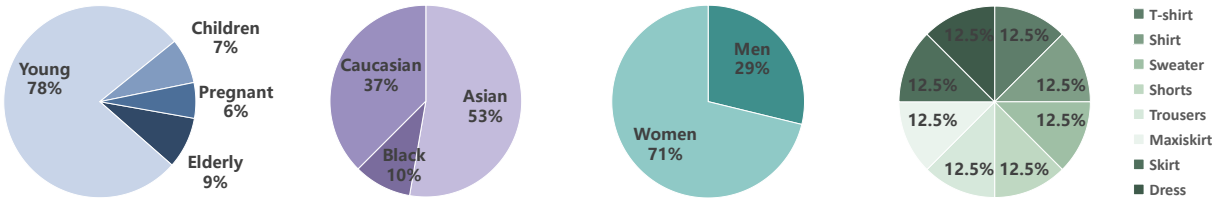
2 Related Work

2.1 Virtual Try-On Methods and Datasets

Recent years have witnessed significant rapid progress in virtual try-on (VTON) research. Representative methods, such as EfficientVITON [1], CatV2TON [7], and StableVITON [18], improve garment detail preservation, generation quality, and inference efficiency. While early studies focused on single-view image-based try-on, recent works [33] have explored multi-view settings and 3D modeling to better capture body–garment interactions. Nevertheless, single-view VTON [6, 7, 13, 18, 19, 26, 28, 31, 41, 46, 47] remains the dominant paradigm and foundation for many recent approaches. Existing VTON datasets are mainly based on VITON-HD [6] and DressCode [29]. VITON-HD provides high-resolution front-view images of upper-body female subjects, while DressCode extends to full-body images, both genders, and diverse garment categories. However, both datasets still lack diversity in poses, body shapes, garment structures, and backgrounds, limiting their ability to represent real-world scenarios and comprehensively evaluate modern VTON models.

2.2 Evaluation of Virtual Try-On Results

Evaluating virtual try-on quality remains challenging. Existing studies mainly rely on objective image similarity metrics, including SSIM [35], LPIPS [49], FID [14], and KID [2], which measure pixel fidelity, perceptual similarity, and distributional realism. However,



(a) Body Shape **(b) Ethnicity** **(c) Gender** **(d) Garment Type**

Figure 4: Distribution of garment type and demographic categories in the dataset.

these metrics often fail to align with human perception in VTON scenarios involving subjective factors such as garment alignment, visual plausibility, and body–cloth interaction.

Recent works have constructed quality assessment datasets across various modalities, including image quality assessment [15, 21, 22, 40], video quality and aesthetics evaluation [16, 23], and task-specific generation scenarios such as digital humans and gesture generation [10, 11, 25, 37, 39]. However, these datasets target generic or other generative tasks and overlook VTON-specific factors. Although some VTON works [6] provide qualitative comparisons, systematic subjective studies with quantitative ratings remain scarce, highlighting the need for perceptually aligned evaluation methods.

3 Dataset and Evaluation Setup

3.1 Dataset construction

For constructing paired data suitable for virtual try-on algorithms, we organize garments into 8 categories with 80 images: **Upper-body**: T-shirt, shirt, sweater; **Lower-body**: shorts, trousers, maxiskirt, skirt; and **Full-body**: dress. The clothing distribution is uniform, as shown in Figure 4(d).

Human subjects are divided into 9 demographic categories, including **Black, Caucasian, Asian, children, young, elderly, pregnant, men, and women**, with proportions shown in Figure 4(a)(b)(c). Beyond these categories, the dataset includes variations in body shapes, proportions, and poses commonly found in virtual try-on scenarios, including standing, hands-on-hips, raised-arm, walking/running, sitting, and waving poses. A total of 184 human images were collected. Representative examples of the eight clothing categories and nine human categories are presented in Figure 3 and Figure 2. The final paired dataset contains 748 garment–person pairs. Since some algorithms cannot process certain images, manual filtering was applied, resulting in 8,132 generated images from various virtual try-on algorithms.

3.2 Virtual try-on algorithms

To generate virtual try-on images, we include a diverse set of representative algorithms spanning traditional and modern paradigms. Specifically, as shown in Table 1, our evaluation covers: **(1) Flow- and warp-based two-stage architectures,** representing the **classical** pipeline of alignment–warping followed by refinement. **(2) Diffusion-based virtual try-on models,** leveraging generative diffusion processes to improve realism and garment fidelity. **(3) Several closed-source commercial or semi-commercial systems,** included to provide references to real-world performance.

Table 1: Categories of virtual try-on algorithms used in our evaluation.

Model	Year	Resolution	Type
VITON-HD [6]	2021	1024x768	Classical (Warp-based)
TPD [46]	2024	384x512	Classical (Warp-based)
DS-VTON [31]	2025	768×1024	Classical (Warp-based)
FS-VTON [13]	2022	256×192	Classical (Warp-based)
Ladi-VTON [28]	2023	1024×768	Diffusion-based
CAT-DM [47]	2024	512×384	Diffusion-based
OOTDiffusion [41]	2024	1024×768	Diffusion-based
StableVITON [18]	2024	1024×768	Diffusion-based
CatV2TON [7]	2025	192×256	Diffusion-based
Kling [19]	2014	512x512-4096x4096	Closed-source
LinkFox [26]	2024	384x384-4096x4096	Closed-source

This comprehensive selection ensures that our dataset and evaluation protocols remain compatible with both earlier pipelines and the latest state-of-the-art virtual try-on techniques.

3.3 Subjective Experiment

After obtaining the virtual try-on results generated by all algorithms, we organized the images into eight non-overlapping groups. A total of 40 independent volunteers with undergraduate and graduate education backgrounds were recruited for subjective evaluation. To facilitate large-scale annotation, each volunteer could participate in the evaluation of one or multiple groups.

The evaluation was conducted along three dimensions, which were scored independently on a 1–5 scale with a precision of 0.1: **(1) Clothing fit:** This metric evaluates whether the target garment is correctly and completely worn in the virtual try-on result. **(2) Body compatibility:** This metric reflects whether the human body shape and pose remain physically consistent after virtual try-on. **(3) Overall quality:** This metric measures whether the final synthesized result aligns with human aesthetic perception. Following data collection, all scores were normalized to ensure consistency across evaluators and to facilitate subsequent statistical analysis and comparison with baseline algorithms. The scoring interface is shown in Figure 5. In addition, prior to the subjective experiments, all participants received standardized training to ensure a clear understanding of the three evaluation dimensions, supported by a dedicated presentation (Figure 6) for detailed explanation.

3.4 Subjective Data Processing

We follow the subjective score processing protocol in [4] for outlier detection and subject reliability screening. For each image, a

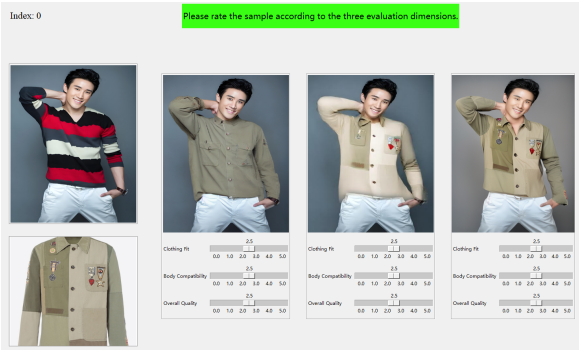


Figure 5: GUI used for subjective quality assessment, where annotators independently rate each virtual try-on result on Clothing Fit, Body Compatibility, and Overall Quality using a 1–5 scale.



Figure 6: Tutorial presentation for participants in the subjective study

rating is an outlier if it deviates from the mean score by more than 2σ (normal distributions) or $\sqrt{20}\sigma$ (non-normal distributions). If over 5% of a subject's ratings are outliers and are approximately symmetrically distributed across score ranges, all ratings from that subject are excluded.

After filtering, the remaining valid scores are normalized using within-subject Z-score normalization and linearly mapped to the range ($[0, 100]$). The Mean Opinion Score (MOS) for each image is then computed by averaging the normalized scores across all valid subjects, formulated as:

$$z_{ij} = \frac{r_{ij} - \mu_i}{\sigma_i}, \quad s_{ij} = \frac{z_{ij} + 3}{6} \times 100, \quad \text{MOS}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} s_{ij} \tag{1}$$

where r_{ij} denotes the raw score given by the i-th subject to the j-th image, μ_i and σ_i represent the mean and standard deviation of all scores provided by subject i, respectively, and N_j is the number of valid ratings for image j.

To evaluate annotation consistency, we perform a partition-based agreement analysis. Since each annotator scores only part of the dataset, annotators within each group are randomly split into two subsets, from which MOS values are computed. SRCC and PLCC between the two MOS sets are calculated over 10 splits and averaged. The results show agreement across all dimensions: Clothing Fit (SRCC/PLCC: 0.7456 ± 0.0253 / 0.7774 ± 0.0195), Body Compatibility (0.7348 ± 0.0295 / 0.7811 ± 0.0341), and Overall Quality (0.7572 ± 0.0273 / 0.7913 ± 0.0306), demonstrating annotation consistency.

3.5 Multi-dimensional Analysis

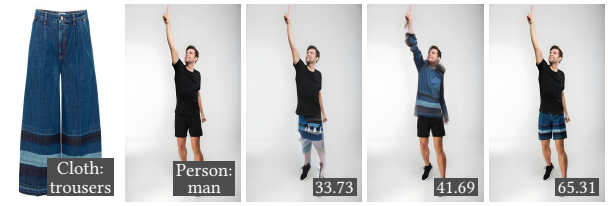
We present a three-dimensional evaluation demo with the corresponding MOS results in Figure 7. The three illustrated score



(a) Clothing fit



(b) Body compatibility



(c) Overall quality

Figure 7: Examples from the proposed VTONQA dataset. We illustrate poor (20–40), average (40–60), and good (60–80) cases for three evaluation dimensions: (a) clothing fit, (b) body compatibility, and (c) overall quality.

ranges, ordered from low to high, reflect performance across different evaluation dimensions: scores below 40 are categorized as poor, scores between 40 and 60 as average, and scores above 60 as good. As shown in Figure 9, performance across the three evaluation dimensions shows consistent trends. Overall, the two closed-source systems outperform all open-source virtual try-on models by a clear margin. Among the remaining algorithms, DS-VTON achieves the best performance within the classical (warp-based) category, while StableVITON leads the diffusion-based methods. Notably, diffusion-based approaches exhibit stronger robustness across diverse scenarios. Although each class of algorithms demonstrates strengths on specific dimensions, performance gaps among open-source methods remain relatively moderate.

The overall distribution of subjective scores is shown in Figure 10. The scores mainly fall within the range of 20–80, where values between 60–80 indicate strong performance and those between 20–40 reflect weaker results. As illustrated by the distribution, most virtual try-on algorithms achieve relatively high scores in body compatibility, suggesting that current models generally preserve human pose without introducing significant structural distortions. In contrast, far fewer methods perform well in clothing fit, that is, accurately aligning the target clothing with the person. This highlights a substantial performance gap when algorithms operate under complex or realistic scenarios. Furthermore, the distribution indicates that overall quality is more heavily influenced by body compatibility than by clothing fit alone.

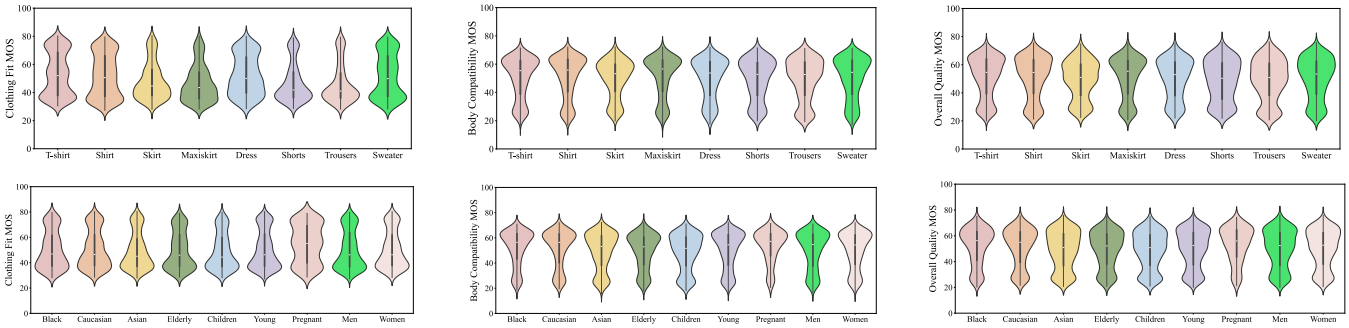


Figure 8: MOS distributions of three evaluation dimensions: (a) garment categories and (b) human body categories.

Table 2: Evaluation of 11 representative VTON models based on the overall quality score. We report both the overall average score and the scores across 8 garment categories and 9 human body categories. ♠ classical (warp-based) method, ♡ diffusion-based method, and ♣ closed-source method. The best results are highlighted in red, the second-best results are highlighted in blue, the third-best results are highlighted in green.

Methods	Clothing Categories								Human Body Categories									Overall
	T-shirt	Shirt	Sweater	Shorts	Trousers	Maxiskirt	Skirt	Dress	Black	Caucasian	Asian	Children	Young	Elderly	Pregnant	Man	Woman	
♠VITON-HD [6]	27.21	25.78	25.77	26.58	25.07	26.28	27.73	25.25	26.22	26.22	26.28	25.80	26.62	26.12	25.73	26.71	26.71	26.26
♠TPD [46]	56.89	54.56	53.78	55.11	56.85	56.62	53.66	56.95	56.10	56.09	54.87	54.74	55.55	54.16	56.84	55.91	55.28	55.48
♠DS-VTON [31]	60.39	61.15	59.28	49.66	49.57	53.86	51.04	54.01	56.19	56.60	55.37	52.83	55.24	54.21	57.64	56.23	54.72	55.64
♠FS-VTON [13]	32.15	30.60	31.58	28.43	25.89	28.17	30.57	28.96	31.09	30.66	29.34	27.66	29.49	29.87	35.39	28.35	29.89	29.79
♡Ladi-VTON [28]	49.81	49.17	47.28	44.80	47.60	52.69	47.45	46.62	47.10	48.89	49.28	45.33	49.07	44.50	48.42	49.05	49.01	48.21
♡CAT-DM [47]	56.03	56.67	55.03	54.86	54.83	59.73	55.08	52.82	58.30	56.30	56.11	52.83	54.42	56.34	55.57	53.96	54.25	55.65
♡OOTDiffusion [41]	37.63	39.16	38.90	36.64	40.38	38.99	38.88	32.94	39.58	38.32	37.31	37.43	37.79	38.70	39.21	37.70	38.02	38.07
♡StableVITON [18]	56.22	56.21	55.44	58.93	56.54	60.18	59.32	52.72	58.00	56.91	57.17	57.14	55.85	56.59	55.67	55.41	56.10	56.68
♡CatV2TON [7]	54.18	53.59	53.48	47.27	46.84	51.15	47.86	52.15	51.80	51.43	51.06	49.07	51.10	50.60	52.46	50.95	51.14	51.21
♣Kling [19]	66.46	67.00	67.87	65.40	66.34	65.50	64.86	66.16	66.00	66.49	66.50	65.90	66.40	64.77	65.36	67.13	66.15	66.35
♣LinkFox [26]	65.56	65.98	64.32	66.47	66.66	64.47	64.17	64.17	64.29	65.02	65.55	64.34	65.69	64.19	64.57	66.04	65.62	65.28

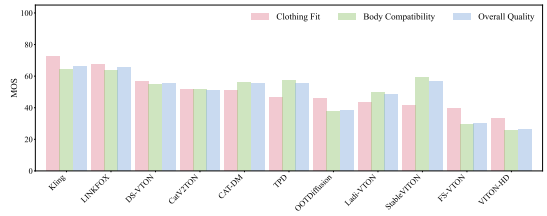


Figure 9: Comparison of the 11 VTON models based on average clothing fit, body compatibility, and overall quality scores.

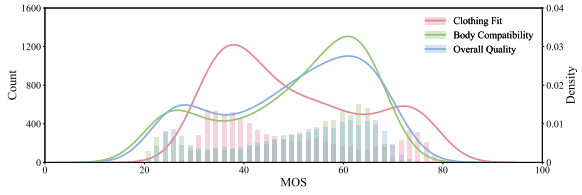


Figure 10: MOS distribution histograms and kernel density curves for clothing fit, body compatibility, and overall quality.

Beyond the overall evaluation, we further examine garment compatibility as the primary dimension of interest, evaluating algorithm performance from two perspectives: score variations across garment categories and across human body categories. The main observations are summarized as follows:

Garment compatibility. As shown in Fig. 8(a), virtual try-on performance for upper-body and full-body garments is consistently and noticeably superior to that for lower-body garments across scenarios. This performance gap is substantial, clearly indicating that current algorithms handle upper-body contours and global garment structures more reliably, while remaining significantly more sensitive to deformation-prone lower-body garments.

Performance across human body categories. As shown in Fig. 8(b), score distributions across human body categories show minor variations, indicating strong generalization. Notably, the pregnant category has a slightly higher proportion of high scores, likely due to dataset construction, as most cases involve upper-body garments where algorithms perform more reliably.

4 Analysis

4.1 Basic Analysis

Table 2 presents processed scores for the Clothing Fit dimension across different garment and human body categories. Beyond the overall comparison, several observations can be drawn. First, diffusion-based methods (e.g., StableVITON) generally outperform classical warp-based approaches (e.g., FS-VTON), while closed-source models consistently achieve the best and most stable performance across categories. Second, garment type plays a more critical role than body category, with lower-body garments (e.g., trousers and shorts) being significantly more challenging than upper-body garments

Table 3: Comparison of IQA metrics on the VTONQA dataset for predicting clothing fit, body compatibility, and overall quality scores. SRCC, KRCC, and PLCC are reported. ♠ traditional full-reference IQA metrics, ♡ traditional no-reference IQA metrics, ♣ deep learning–based full-reference IQA methods, and ♢ deep learning–based no-reference IQA methods. Fine-tuned results are marked with *. The best results are highlighted in red, and the second-best results are highlighted in blue.

Methods	Clothing Fit			Body Compatibility			Overall Quality		
	SRCC↑	KRCC↑	PLCC↑	SRCC↑	KRCC↑	PLCC↑	SRCC↑	KRCC↑	PLCC↑
♠MSE	-0.0350	-0.0258	-0.0289	0.3117	-0.2189	-0.2911	0.2688	0.1886	0.2483
♠PSNR	-0.0350	-0.0258	-0.1009	0.3117	0.2189	0.3592	0.2688	0.1886	0.3126
♠SSIM [35]	0.0563	0.0376	0.0794	0.3299	0.2250	0.3302	0.2912	0.1982	0.2946
♠FSIM [48]	0.0799	0.0516	0.0478	0.4079	0.2848	0.4390	0.3743	0.2610	0.3991
♠SCSSIM [12]	0.0394	0.0270	0.0667	0.3164	0.2159	0.3147	0.2767	0.1888	0.2785
♠GMSD [43]	0.1083	0.0721	0.1168	0.1970	0.1325	0.1917	0.1947	0.1315	0.1986
♡BRISQUE [27]	0.1009	0.0654	0.1428	0.1778	0.1181	0.1679	0.1732	0.1151	0.1735
♣LPIPS(alex) [49]	0.0615	0.0360	0.0831	0.4292	0.3024	0.4967	0.3918	0.2761	0.4532
♣LPIPS(vgg) [49]	0.1395	0.0872	0.1920	0.4927	0.3473	0.5522	0.4572	0.3214	0.5160
♣AHIQ [20]	-0.0716	-0.0505	-0.0676	0.2173	0.1478	0.2605	0.1773	0.1204	0.2147
♢CNNIQA [17]	0.0326	0.0231	-0.1102	0.1126	0.0793	0.0703	0.1128	0.0801	0.0446
♢WaDIQaM [3]	-0.0030	0.0011	-0.1399	0.0872	0.0656	0.0813	0.0789	0.0606	0.0493
♢NIMA [32]	0.3194	0.2159	0.2969	0.4322	0.2966	0.5092	0.4318	0.2953	0.4969
♢HyperIQA [30]	0.1335	0.0759	0.1120	0.2876	0.1834	0.4305	0.2785	0.1762	0.3984
♢TOPIQ* [5]	0.3157	0.2156	0.3010	0.3666	0.2767	0.4586	0.3930	0.2745	0.4466
♢MANIQA* [44]	0.6733	0.4813	0.6334	0.6654	0.4813	0.8011	0.7072	0.5119	0.7967
♢CLIPIQA* [34]	0.4420	0.3108	0.4190	0.3010	0.2176	0.4550	0.3720	0.2658	0.5000

(e.g., T-shirt and shirt). Third, some open-source methods (e.g., DS-VTON and StableVITON) demonstrate competitive performance in specific categories but lack consistency across all settings, indicating a trade-off between specialization and generalization. Finally, the large performance gap among methods highlights the substantial room for improvement in current virtual try-on approaches.

4.2 Baseline Experiment

We evaluate a diverse set of baseline methods, including: (1) traditional full-reference (FR) IQA metrics: MSE, PSNR, SSIM [35], FSIM [48], SCSSIM [12], GMSD [43]; (2) traditional no-reference (NR) IQA metrics: BRISQUE [27], (3) deep learning–based FR IQA methods: LPIPS(alex) [49], LPIPS(vgg) [49], AHIQ [20]; (4) deep learning–based NR IQA methods: CNNIQA [17], WaDIQaM [3], NIMA [32], HyperIQA [30], TOPIQ [5], MANIQA [44], CLIPIQA [34]. Among the deep learning–based NR IQA approaches, several models are further fine-tuned using the proposed subjective dataset. For models without additional training, the virtual try-on images are directly fed into the baseline algorithms to assess which of the three perceptual dimensions the models naturally align with. For models that are further fine-tuned, each dimension's scores are individually used to train the model, verifying how our dataset can improve dimension-specific evaluation performance.

We use Spearman's Rank Correlation Coefficient(SRCC), Pearson Linear Correlation Coefficient(PLCC), and Kendall's Rank Correlation Coefficient(KRCC) to evaluate the consistency between predicted quality scores and ground-truth MOSs. The resulting values are reported in Table 3.

From the table, several observations emerge: First, traditional IQA metrics—both full-reference (FR) and no-reference (NR)—correlate poorly with human perception of virtual try-on results. Pixel-level similarity metrics are inconsistent, failing to capture garment deformation and body–cloth interactions, while conventional NR-IQA methods based on natural image statistics are limited in this scenario. Second, perceptual distance–based metrics perform better on body compatibility and overall quality but struggle with clothing fit. Third, deep learning–based NR-IQA methods outperform traditional metrics, with transformer- or multi-modal approaches (e.g., MANIQA) achieving the best results. Notably, clothing fit remains the most challenging dimension.

4.3 Future work

Although VTONQA provides a comprehensive benchmark for perceptual quality assessment of virtual try-on results, several limitations remain. The current dataset mainly focuses on standard single-garment scenarios, with limited coverage of layered garments, and does not provide dense human annotations (e.g., human parsing maps and pose labels) for fine-grained analysis. Future work will expand the dataset by increasing garment diversity and human subjects, incorporating more complex virtual try-on scenarios, and introducing fine-grained distortion annotations to enable more precise diagnosis of garment deformation, local misalignment, and visual artifacts. In addition, we plan to further standardize image quality and develop a more structured evaluation framework for next-generation virtual try-on models.

5 Conclusion

In this work, we present VTONQA, the first quality assessment dataset for VTON, comprising 8,132 VTON-generated images and 24,396 MOS annotations across three dimensions: clothing fit, body compatibility, and overall quality. Through subjective assessment of representative VTON methods, we identify key factors affecting try-on quality and reveal limitations of existing objective metrics, emphasizing subjective supervision. We believe VTONQA and its evaluation framework will facilitate perceptually aligned quality assessment methods and reliable VTON algorithms.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62225112, 62401365, and in part by the China Postdoctoral Science Foundation under Grant Number BX20250411, 2025M773473.

References

[1] Mostafa Atef, Mariam Ayman, Ahmed Rashed, Ashrakat Saeed, Abdelrahman Saeed, and Ahmed Fares. 2025. EfficientVITON: An Efficient Virtual Try-On Model using Optimized Diffusion Process. *arXiv preprint arXiv:2501.11776* (2025).

[2] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. 2018. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401* (2018).

[3] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. 2018. Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment. *IEEE Transactions on Image Processing* 27, 1 (2018), 206–219. doi:10.1109/TIP.2017.2760518

[4] RIR BT. 2002. Methodology for the subjective assessment of the quality of television pictures. *International Telecommunication Union* 4 (2002), 19.

[5] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2024. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing* 33 (2024), 2404–2418.

[6] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. 2021. VITON-HD: High-Resolution Virtual Try-On via Misalignment-Aware Normalization. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*.

[7] Zheng Chong, Wenqing Zhang, Shiyue Zhang, Jun Zheng, Xiao Dong, Haoxiang Li, Yiling Wu, Dongmei Jiang, and Xiaodan Liang. 2025. Catv2ton: Taming diffusion transformers for vision-based virtual try-on with temporal concatenation. *arXiv preprint arXiv:2501.11325* (2025).

[8] Huiyu Duan, Kang Fu, Sijing Wu, Yunhao Li, Zicheng Zhang, Qiang Hu, Xiongkuo Min, and Guangtao Zhai. 2025. BMPCQA: Bioinspired Metaverse Point Cloud Quality Assessment Based on Large Multimodal Models. *Advanced Intelligent Systems* (2025), 2500504.

[9] Yixuan Gao, Xiongkuo Min, Jinliang Han, Yuqin Cao, Sijing Wu, Yunze Dou, and Guangtao Zhai. 2025. Multi-Dimensional Text-to-Face Image Quality Assessment Using LLM: Database and Method. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 6948–6957.

[10] Zhilin Gao, Yunhao Li, Sijing Wu, Yuqin Cao, Huiyu Duan, and Guangtao Zhai. 2025. Ges-qa: A multidimensional quality assessment dataset for audio-to-3d gesture generation. *arXiv preprint arXiv:2508.12020* (2025).

[11] Zhilin Gao, Yunhao Li, Sijing Wu, Yucheng Zhu, Huiyu Duan, and Guangtao Zhai. 2026. SFQA: A Comprehensive Perceptual Quality Assessment Dataset for Singing Face Generation. *arXiv preprint arXiv:2601.20385* (2026).

[12] Ke Gu, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang. 2012. An improved full-reference image quality metric based on structure compensation. In *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*. 1–6.

[13] Sen He, Yi-Zhe Song, and Tao Xiang. 2022. Style-Based Global Appearance Flow for Virtual Try-On. In *CVPR*.

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

[15] Yanwei Jiang, Wei Sun, Yingjie Zhou, Xiangyang Zhu, Yuqin Cao, Jun Jia, Yunhao Li, Sijing Wu, Dandan Zhu, Xingkuo Min, et al. 2026. Surveillance Facial Image Quality Assessment: A Multi-dimensional Dataset and Lightweight Model. *IEEE Transactions on Circuits and Systems for Video Technology* (2026).

[16] Jianing Jin, Jiangyong Ying, Huiyu Duan, Liu Yang, Sijing Wu, Yunhao Li, Yushuo Zheng, Xiongkuo Min, and Guangtao Zhai. 2025. RGC-VQA: An Exploration Database for Robotic-Generated Video Quality Assessment. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 13413–13420.

[17] Le Kang, Peng Ye, Yi Li, and David Doermann. 2014. Convolutional Neural Networks for No-Reference Image Quality Assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[18] Jeongho Kim, Guojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. 2024. 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*. 8176–8185.

[19] Kuaishou Technology. 2024. *Kling AI Virtual Try-On*. https://app.klingai.com/cn/try-on/try-on/new

[20] Shanshan Lao, Yuan Gong, Shuwei Shi, Sidi Yang, Tianhe Wu, Jiahao Wang, Weihao Xia, and Yujiu Yang. 2022. Attentions help cnns see better: Attention-based hybrid image quality assessment network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1140–1149.

[21] Xinyue Li, Zhiming Xu, Zhichao Zhang, Zhaolin Cai, Sijing Wu, Xiongkuo Min, Yitong Chen, and Guangtao Zhai. 2026. ELIQ: A Label-Free Framework for Quality Assessment of Evolving AI-Generated Images. *arXiv preprint arXiv:2602.03558* (2026).

[22] Yunhao Li, Sijing Wu, Huiyu Duan, Yucheng Zhu, Qi Jia, and Guangtao Zhai. 2025. Exploring Instruction Data Quality for Explainable Image Quality Assessment. *arXiv preprint arXiv:2510.03880* (2025).

[23] Yunhao Li, Sijing Wu, Zhilin Gao, Zicheng Zhang, Qi Jia, Huiyu Duan, Xiongkuo Min, and Guangtao Zhai. 2026. VideoAesBench: Benchmarking the Video Aesthetics Perception Capabilities of Large Multimodal Models. *arXiv preprint arXiv:2601.21915* (2026).

[24] Yunhao Li, Sijing Wu, Wei Sun, Zhichao Zhang, Yucheng Zhu, Zicheng Zhang, Huiyu Duan, Xiongkuo Min, and Guangtao Zhai. 2025. AGHI-QA: A Subjective-Aligned Dataset and Metric for AI-Generated Human Images. *arXiv preprint arXiv:2504.21308* (2025).

[25] Yunhao Li, Sijing Wu, Yucheng Zhu, Huiyu Duan, Zicheng Zhang, and Guangtao Zhai. 2025. DHQA-4D: Perceptual Quality Assessment of Dynamic 4D Digital Human. *arXiv preprint arXiv:2510.03874* (2025).

[26] LinkFox. 2024. *LinkFox AI Dressing: Virtual Try-On Platform*. https://www.linkfox.com/draw/ai-dressing

[27] Anish Mittal, Anush K Moorthy, and Alan C Bovik. 2011. Blind/referenceless image spatial quality evaluator. In *2011 conference record of the forty fifth asilomar conference on signals, systems and computers (ASILOMAR)*. IEEE, 723–727.

[28] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. 2023. LaDI-VTON: Latent Diffusion Textual-Inversion Enhanced Virtual Try-On. In *Proceedings of the ACM International Conference on Multimedia*.

[29] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. 2022. Dress Code: High-Resolution Multi-Category Virtual Try-On. In *Proceedings of the European Conference on Computer Vision*.

[30] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. 2020. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3667–3676.

[31] Xianbing Sun, Yan Hong, Jiahui Zhan, Jun Lan, Huijia Zhu, Weiqiang Wang, Liqing Zhang, and Jianfu Zhang. 2025. DS-VTON: High-Quality Virtual Try-on via Disentangled Dual-Scale Generation. *arXiv preprint arXiv:2506.00908* (2025).

[32] Hossein Talebi and Peyman Milanfar. 2018. NIMA: Neural Image Assessment. *IEEE Transactions on Image Processing* 27, 8 (2018), 3998–4011. doi:10.1109/TIP.2018.2831899

[33] Haoyu Wang, Zhilu Zhang, Donglin Di, Shiliang Zhang, and Wangmeng Zuo. 2024. MV-VTON: Multi-View Virtual Try-On with Diffusion Models. *arXiv preprint arXiv:2404.17364* (2024).

[34] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. 2023. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 2555–2563.

[35] Z. WANG. 2004. Image quality assessment : Form error visibility to structural similarity. *IEEE Trans. Image Process.* 13, 4 (2004), 604–606. https://cir.nii.ac.jp/crid/1570009751473985280

[36] Sijing Wu, Yunhao Li, Huiyu Duan, Yanwei Jiang, Yucheng Zhu, and Guangtao Zhai. 2025. HVEval: Towards Unified Evaluation of Human-Centric Video Generation and Understanding. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 13376–13383.

[37] Sijing Wu, Yunhao Li, Huiyu Duan, Yucheng Zhu, Xiongkuo Min, Patrick Le Callet, and Guangtao Zhai. 2026. Multi-Dimensional Quality Assessment for AI-Generated Human-Centric Videos: Dataset and Model. *IEEE Transactions on Circuits and Systems for Video Technology* (2026).

[38] Sijing Wu, Yunhao Li, Ziwen Xu, Yixuan Gao, Huiyu Duan, Wei Sun, and Guangtao Zhai. 2025. FVQ: A Large-Scale Dataset and an LMM-based Method for Face Video Quality Assessment. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 6928–6937.

[39] Sijing Wu, Yunhao Li, Weitian Zhang, Jun Jia, Yucheng Zhu, Yichao Yan, Guangtao Zhai, and Xiaokang Yang. 2025. Singinghead: A Large-Scale 4D Dataset for Singing Head Animation. *IEEE Transactions on Multimedia* (2025).

[40] Sijing Wu, Yunhao Li, Zicheng Zhang, Qi Jia, Xinyue Li, Huiyu Duan, Xiongkuo Min, and Guangtao Zhai. 2026. Q-Bench-Portrait: Benchmarking Multimodal Large Language Models on Portrait Image Quality Perception. *arXiv preprint arXiv:2601.18346* (2026).

[41] Yuhao Xu, Tao Gu, Weifeng Chen, and Chengcai Chen. 2024. OOTDiffusion: Outfitting Fusion based Latent Diffusion for Controllable Virtual Try-on. *arXiv preprint arXiv:2403.01779* (2024).

[42] Zitong Xu, Huiyu Duan, Bingnan Liu, Guangji Ma, Jiarui Wang, Liu Yang, Shiqi Gao, Xiaoyu Wang, Jia Wang, Xiongkuo Min, et al. 2025. Lmm4edit: Benchmarking and evaluating multimodal image editing with lmms. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 6908–6917.

[43] Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C. Bovik. 2014. Gradient Magnitude Similarity Deviation: A Highly Efficient Perceptual Image Quality Index. *IEEE Transactions on Image Processing* 23, 2 (2014), 684–695. doi:10.1109/TIP.2013.2293423

[44] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. 2022. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1191–1200.

[45] Woo Yi Yang, Jiarui Wang, Sijing Wu, Huiyu Duan, Yuxin Zhu, Liu Yang, Kang Fu, Guangtao Zhai, and Xiongkuo Min. 2025. Lmme3dhf: Benchmarking and evaluating multimodal 3d human face generation with lmms. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 8825–8834.

[46] Xu Yang, Changxing Ding, Zhibin Hong, Junhao Huang, Jin Tao, and Xiangmin Xu. 2024. Texture-Preserving Diffusion Models for High-Fidelity Virtual Try-On. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7017–7026.

[47] Jianhao Zeng, Dan Song, Weizhi Nie, Hongshuo Tian, Tongtong Wang, and An-An Liu. 2024. CAT-DM: Controllable Accelerated Virtual Try-on with Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8372–8382.

[48] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. 2011. FSIM: A Feature Similarity Index for Image Quality Assessment. *IEEE Transactions on Image Processing* 20, 8 (2011), 2378–2386. doi:10.1109/TIP.2011.2109730

[49] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[50] Yingjie Zhou, Zicheng Zhang, Sijing Wu, Jun Jia, Yanwei Jiang, Wei Sun, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. 2026. MI3S: A multimodal large language model assisted quality assessment framework for AI-generated talking heads. *Information Processing & Management* 63, 1 (2026), 104321.