

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

Xinyi Wei, Sijing Wu, Zitong Xu, Yunhao Li, Huiyu Duan, Xiongkuo Min, Guangtao Zhai
 Shanghai Jiao Tong University
 {moj-will, wusijing, xuzitong, lyhsjtu, huiyuduan, minxiongkuo, zhaiguangta}@sjtu.edu.cn

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.

Index Terms—Virtual try-on, dataset, benchmark, subjective experiment, large multi-modal models

I. INTRODUCTION

Image-based virtual try-on (VTON) [1]–[3] has emerged as a promising technology that enables realistic visualization of garments on human images (see Figure 1), with broad applications in e-commerce, virtual reality, and digital fashion. However, current VTON models often struggle to generate high-quality results, exhibiting artifacts such as blurred faces and garments, distorted body structures, and failures in proper clothing transfer, which severely degrade user experience and practical applicability. Therefore, effective evaluation of VTON-generated images is crucial for monitoring perceptual quality in real-world VTON applications, 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) [4], the perceptual similarity metric Learned Perceptual Image Patch Similarity (LPIPS) [5], and traditional pixel-level criteria such as Structural Similarity Index (SSIM) [6] 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 [7]–[15] typically learn a network to regress quality scores based on human-annotated datasets and are inherently aligned with human perception. However, existing QA datasets are primarily designed for natural images or specific Artificial Intelligence (AI)-generated images, and

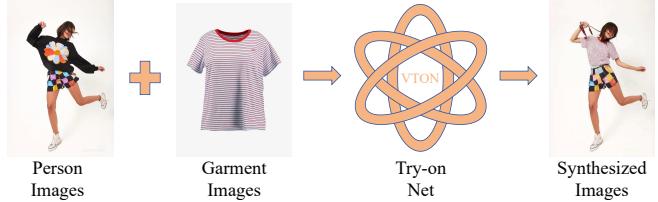


Fig. 1: Illustration of the image-based virtual try-on pipeline.

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 [12], [13], [16], QA models trained on existing datasets are often inapplicable to or perform poorly on VTON-generated images, highlighting the urgent need for a QA dataset specifically designed for VTON.

To bridge this gap, we construct **VTONQA**, the first large-scale multi-dimensional quality assessment dataset for VTON-generated images, comprising 8,132 images from 11 representative VTON models and 24,396 mean opinion scores (MOSS) across three evaluation dimensions: clothing fit, body compatibility, and overall quality. Specifically, the VTON-generated images are synthesized by applying try-on garments from 8 categories to 183 reference person images spanning 9 categories. The VTON models include classical warp-based [1], [3], [17], [18], diffusion-based [2], [19]–[22], and closed-source methods [23], [24]. Subsequently, 40 subjects are recruited to annotate the images across the three evaluation dimensions, under the supervision of a professional team of image processing researchers to ensure annotation quality. Based on the VTONQA dataset, we benchmark the try-on capabilities of 11 VTON models and the quality assessment capabilities of 17 image quality assessment (IQA) metrics. All VTON models are evaluated in an inference-only setting, without any additional fine-tuning or retraining on VTONQA. For quality assessment, we include both full-reference and no-reference IQA metrics, spanning traditional and deep learning-based methods, and show that fine-tuned models achieve higher correlation with human perceptual judgments, highlighting the significance of the proposed VTONQA dataset.

We hope that the proposed VTONQA dataset, together with the provided benchmarks, will foster in-depth research on objective quality assessment methods for VTON-generated images that better align with human perception, ultimately

advancing the development of VTON models.

In summary, the main contributions of this work are:

- To the best of our knowledge, we are the first to conduct a comprehensive subjective quality assessment study of VTON-generated images.
- We build the first multi-dimensional quality assessment dataset for VTON-generated images, which comprises 8,132 images and 24,396 MOS annotations across three dimensions (*i.e.*, clothing fit, body compatibility, and overall quality).
- Based on the VTONQA dataset, we benchmark the performance of 11 VTON models and the quality assessment capabilities of 17 IQA metrics.

II. RELATED WORK

A. Virtual Try-On Methods and Datasets

Recent years have witnessed rapid progress in virtual try-on (VTON) research. Representative methods such as EfficientVTON [25], CatV2TON [22], and StableVTON [2] focus on improving garment detail preservation, generation quality, and inference efficiency. While early studies mainly addressed single-view image-based try-on, subsequent works [26] have extended toward multi-view settings and 3D modeling to better capture body–garment interactions. Nevertheless, single-view VTON [1]–[3], [17]–[24] remains the most mature and widely adopted paradigm, forming the basis of many recent approaches. Existing VTON datasets are dominated by VITON-HD [1] and DressCode [27]. VITON-HD provides high-resolution front-view images of upper-body female subjects, whereas DressCode extends to full-body images, both genders, and multiple garment categories. Despite their scale and resolution, these datasets exhibit limited diversity in pose, body shape, garment structure, and background complexity, which restricts their ability to reflect real-world scenarios and to comprehensively evaluate modern VTON models.

B. Evaluation of Virtual Try-On Results

Evaluating the quality of virtual try-on results remains challenging. Most existing studies rely on objective image similarity metrics, including SSIM [6], LPIPS [5], FID [4], and KID [28], which measure pixel-level fidelity, perceptual similarity, and distributional realism. These metrics are efficient and widely adopted as standard evaluation tools.

However, objective metrics often fail to align with human perceptual judgments, particularly in virtual try-on scenarios where garment alignment, visual plausibility, and body–cloth interaction are highly subjective. Although some works [1] provide qualitative visual comparisons, systematic subjective studies with quantitative human ratings remain scarce. This mismatch between objective metrics and human perception highlights the necessity of incorporating subjective evaluation to achieve more reliable and perceptually meaningful assessment.

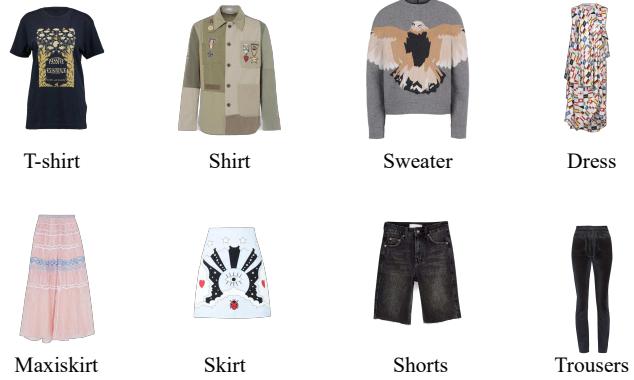


Fig. 2: Examples for each clothing category in VTONQA.

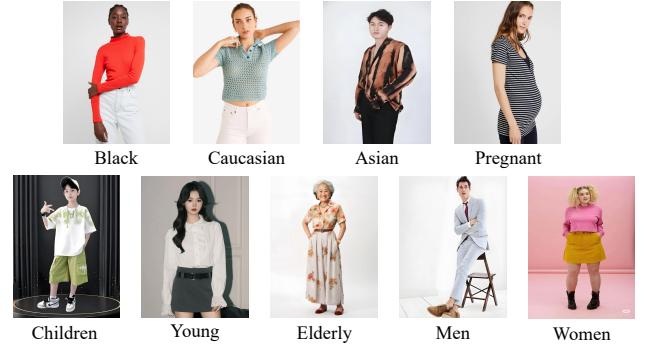


Fig. 3: Examples for each human body category in VTONQA.

III. DATASET AND EVALUATION SETUP

A. Dataset construction

For constructing paired data suitable for virtual try-on algorithms, we organize garments into 8 categories, with a total of 80 images: **Upper-body**: T-shirt, shirt, sweater; **Lower-body**: shorts, trousers, maxiskirt, skirt; **Full-body**: dress.

Human subjects are grouped into 9 demographic categories, including **Black**, **Caucasian**, **Asian**, **children**, **young**, **elderly**, **pregnant**, **men**, and **women**. In addition to these major categories, the dataset also covers finer-grained variations—such as different body shapes and body proportions—which are present in the data but not explicitly used as classification labels during dataset construction. This design allows the dataset to maintain structured organization while still capturing the natural diversity of real-world human appearances. A total of 189 human images were collected. Figure 2 and Figure 3 show the dataset composition.

The final paired dataset comprises 748 garment–person pairs, resulting in 8,132 images generated through various virtual try-on algorithms.

B. 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 I, our evaluation covers: **(1) Flow-based and warp-based two-stage architectures**, which represent the **classical** pipeline of alignment–warping followed by refinement. **(2) Diffusion-based**

Model	Year	Resolution	Type
VITON-HD [1]	2021	1024x768	Classical (Warp-based)
TPD [3]	2024	384x512	Classical (Warp-based)
DS-VTON [17]	2025	768x1024	Classical (Warp-based)
FS-VTON [18]	2022	256x192	Classical (Warp-based)
Ladi-VTON [19]	2023	1024x768	Diffusion-based
CAT-DM [20]	2024	512x384	Diffusion-based
OOTDiffusion [21]	2024	1024x768	Diffusion-based
StableVTON [2]	2024	1024x768	Diffusion-based
CatV2TON [22]	2025	192x256	Diffusion-based
Kling [23]	2014	512x512-4096x4096	Closed-source
LinkFox [24]	2024	384x384-4096x4096	Closed-source

TABLE I: Categories of virtual try-on algorithms used in our evaluation.



Fig. 4: Illustration of the GUI used in the subjective study.

virtual try-on models, which leverage generative diffusion processes to improve realism and garment fidelity. **(3) Several closed-source commercial or semi-commercial systems**, included to provide additional references to real-world performance.

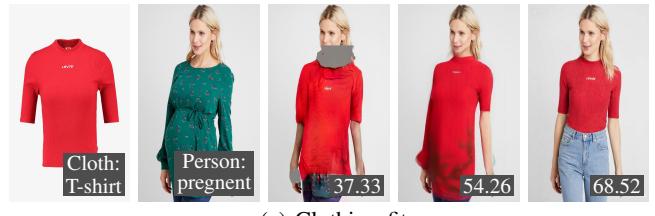
This comprehensive selection ensures that our dataset and evaluation protocols remain compatible with both earlier pipelines and the latest state-of-the-art virtual try-on techniques.

C. Subjective Experiment

After obtaining the virtual try-on results generated by all algorithms, we organized the images into eight groups. For each group, five independent volunteers were recruited to provide subjective ratings. The evaluation was conducted along three dimensions:

(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.

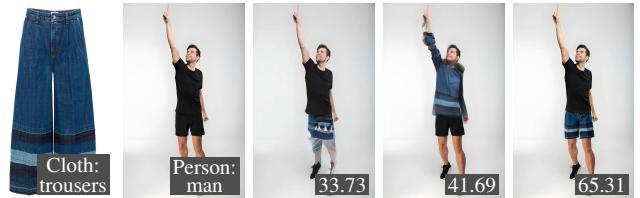
The detailed scoring criteria are summarized in Figure 5. Following data collection, all scores were normalized to ensure consistency across evaluators and to facilitate subsequent statistical analysis and comparison with baseline algorithms. And the scoring interface is shown in the figure 4.



(a) Clothing fit



(b) Body compatibility



(c) Overall quality

Fig. 5: 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.

D. Subjective Data Processing

We follow the subjective score processing protocol proposed in [29] to perform outlier detection and subject reliability screening. For each image sample, an individual rating is considered an outlier if it deviates from the mean score of that image by more than (2σ) (for approximately normal score distributions) or $(\sqrt{20}\sigma)$ (for non-normal distributions). Furthermore, if more than 5% of a subject's ratings are identified as outliers and these outliers are approximately symmetrically distributed across high and low 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:

$$\text{MOS}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} \frac{(r_{ij} - \mu_i)/\sigma_i + 3}{6} \times 100 \quad (1)$$

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 .

E. Multi-dimensional Analysis

As shown in Figure 7, the performance across the three evaluation dimensions exhibits consistent patterns. Overall,

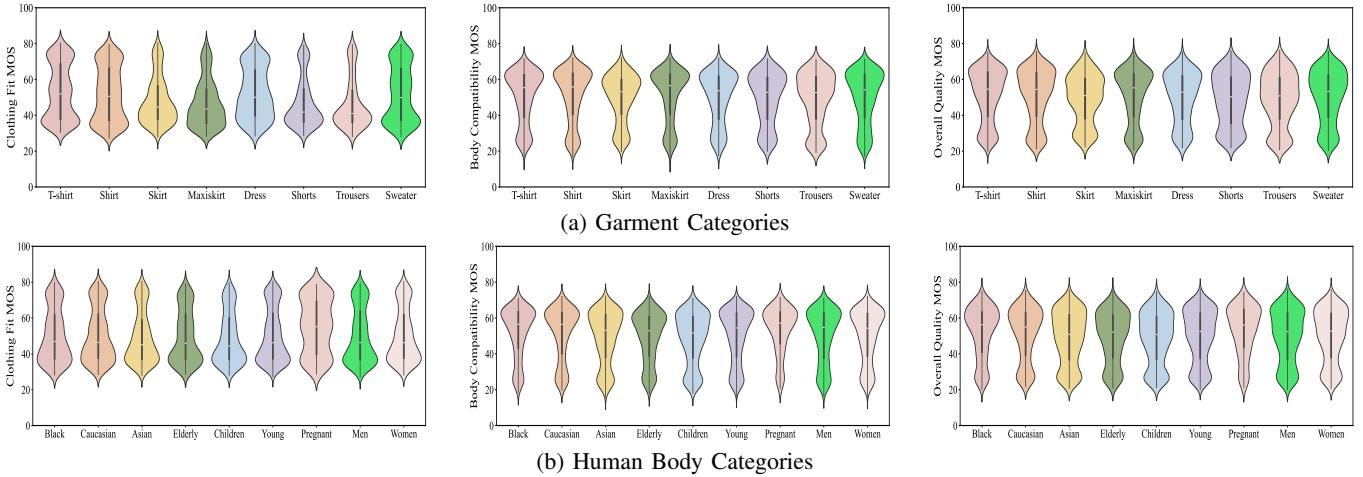


Fig. 6: MOS distributions of the three evaluation dimensions (clothing fit, body compatibility, and overall quality) across (a) eight garment categories and (b) nine human body categories, respectively.

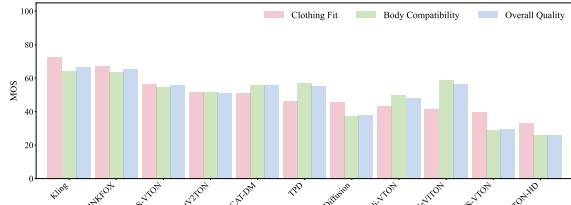


Fig. 7: Comparison of the 11 VTON models based on average clothing fit, body compatibility, and overall quality scores.

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, and StableVITON leads the diffusion-based methods. While each class of algorithms demonstrates strengths on specific dimensions, the performance gaps among open-source methods remain relatively moderate.

The overall distribution of subjective scores is shown in Figure 8. The scores primarily 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 deviations. In contrast, far fewer methods perform well in clothing fit, that is, accurately fitting the target clothing onto the person. This highlights a substantial performance gap when algorithms operate under complex or realistic scenarios. Furthermore, the distribution indicates that overall quality are more heavily influenced by body compatibility than by clothing fit alone.

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:

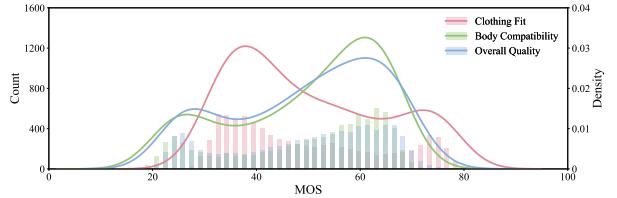


Fig. 8: MOS distribution histograms and kernel density curves for clothing fit, body compatibility, and overall quality.

Garment compatibility. As shown in Fig. 6(a), virtual try-on performance for upper-body and full-body garments is consistently superior to that for lower-body garments. This performance gap is substantial, indicating that current algorithms handle upper-body contours and global garment structures more reliably, while remaining more sensitive to deformation-prone lower-body garments.

Performance across human body categories. As illustrated in Fig. 6(b), the score distributions across different human body categories exhibit only minor variations, suggesting that existing virtual try-on algorithms maintain relatively strong generalization across diverse body types. Notably, the pregnant category shows a slightly higher proportion of high scores compared to other categories. This trend is likely attributable to the dataset construction, as most evaluated pregnant subjects involve upper-body garments only, for which the majority of algorithms demonstrate stronger and more stable performance.

IV. ANALYSIS

A. Basic Analysis

The table II the processed scores for one evaluation dimension (Clothing Fit) across different garment categories and human body categories. Clearly, the closed-source algorithms outperform others across all aspects. Notably, among the open-source methods, DS-VTON and StableVITON demonstrates comparable performance to closed-source algorithms in specific areas, such as upper-body garments.

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 [1]	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 [3]	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 [17]	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 [18]	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 [19]	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 [20]	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 [21]	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
♡StableVTON [2]	56.22	56.21	55.44	58.93	56.54	60.18	59.32	52.72	58.00	56.91	57.17	55.85	56.59	55.67	55.41	56.10	56.68	
♡CatVTON [22]	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 [23]	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 [24]	65.56	65.98	64.32	66.47	66.66	64.47	64.17	64.29	65.02	65.55	64.34	65.69	64.19	64.57	66.04	65.62	65.28	

TABLE II: 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 Fit			Body Compatibility			Overall Quality			
	ρ_s	ρ_k	ρ_p	ρ_s	ρ_k	ρ_p	ρ_s	ρ_k	ρ_p	
♠MSE	-0.035	-0.026	-0.029	0.312	-0.219	-0.291	0.269	0.189	0.248	
♠PSNR	-0.035	-0.026	-0.101	0.312	0.219	0.359	0.269	0.189	0.313	
♠SSIM [6]	0.056	0.038	0.079	0.330	0.225	0.330	0.291	0.198	0.295	
♠FSIM [30]	0.080	0.052	0.048	0.408	0.285	0.439	0.374	0.261	0.399	
♠SCSSSIM [31]	0.039	0.027	0.067	0.316	0.216	0.315	0.277	0.189	0.279	
♠GMSD [32]	0.108	0.072	0.117	0.197	0.133	0.192	0.195	0.132	0.199	
♡BRISQUE [33]	0.101	0.065	0.143	0.178	0.118	0.168	0.173	0.115	0.174	
♣LPIPS(alex) [5]	0.062	0.036	0.083	0.429	0.302	0.497	0.392	0.276	0.453	
♣LPIPS(vgg) [5]	0.140	0.087	0.192	0.493	0.347	0.552	0.457	0.321	0.516	
♣AHIQ [34]	-0.072	-0.051	-0.068	0.217	0.148	0.261	0.177	0.120	0.215	
◊CNNIQA [35]	0.033	0.023	-0.110	0.113	0.079	0.070	0.113	0.080	0.045	
◊WaDIQaM [36]	-0.003	0.001	-0.140	0.087	0.066	0.813	0.079	0.061	0.049	
◊NIMA [37]	0.319	0.216	0.297	0.432	0.297	0.509	0.432	0.295	0.497	
◊HyperIQA [38]	0.134	0.076	0.112	0.288	0.183	0.431	0.279	0.176	0.398	
◊TOPIO* [39]	0.291	0.194	0.222	0.367	0.258	0.503	0.393	0.275	0.447	
◊MANIQA* [7]	0.673	0.481	0.633	0.665	0.481	0.801	0.707	0.512	0.797	
◊CLPIQA* [8]	0.442	0.311	0.419	0.301	0.218	0.455	0.372	0.266		

TABLE III: Comparison of IQA metrics on the VTONQA dataset for predicting clothing fit, body compatibility, and overall quality scores. SRCC (ρ_s), KRCC (ρ_k), and PLCC (ρ_p) 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.

B. Baseline Experiment

We evaluate a diverse set of baseline methods, including: traditional full-reference (FR) IQA metrics, traditional no-reference (NR) IQA metrics, deep learning-based FR IQA methods, and deep learning-based NR IQA methods. Among the deep learning-based NR IQA approaches, several models are further fine-tuned using the proposed subjective dataset. The resulting values are reported in Table III.

From the table, several observations can be made:

First, traditional IQA metrics—including both full-reference (FR) and no-reference (NR) methods—generally show low correlation with human perception of virtual try-on results. Pixel-level similarity metrics are particularly inconsistent, failing to capture perceptual effects caused by garment deformation and body–garment interactions, while conventional NR-IQA methods based on natural image statistics have limited modeling capability for this scenario.

Second, perceptual distance-based metrics achieve better performance on body compatibility and overall quality but

remain less effective for accurately assessing clothing fit.

Third, deep learning-based NR-IQA methods substantially outperform traditional metrics, with transformer- or multimodal feature-based approaches (e.g., MANIQA) achieving the best results. Notably, evaluation difficulty varies across perceptual dimensions, with clothing fit emerging as the most challenging to predict reliably.

C. Future work

Considering the current limitations of the dataset, future work will focus on expanding the dataset and enriching the evaluation framework.

Dataset Expansion: Although the dataset covers diverse garment–person pairs, samples per category are still limited. Future work will expand the dataset by increasing garment diversity and the number of human subjects for more comprehensive pairing. While prioritizing realism, the dataset contains images with varying quality (e.g., resolution and lighting). As it is designed primarily for evaluation, future work will standardize quality and reorganize the dataset into a structured form suitable for next-generation virtual try-on models.

Enhanced Evaluation Framework: Beyond global subjective scores, future extensions will incorporate fine-grained distortion annotations to highlight local misalignment, garment deformation, and visual artifacts, enabling more precise diagnosis of algorithm weaknesses and targeted improvements for virtual try-on models.

V. CONCLUSION

In this work, we present the first quality assessment dataset for VTON, termed VTONQA, which comprises 8,132 VTON-generated images and 24,396 MOS annotations across three perceptual dimensions, namely clothing fit, body compatibility, and overall quality. Through comprehensive multi-dimensional subjective assessments of representative VTON methods, we identify key factors affecting try-on quality and expose the limitations of existing objective metrics, highlighting the importance of subjective supervision. We believe that the proposed dataset and evaluation framework will facilitate the development of more perceptually aligned quality assessment methods and more reliable VTON algorithms.

REFERENCES

[1] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo, “Viton-hd: High-resolution virtual try-on via misalignment-aware normalization,” in *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2021.

[2] 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*, 2024, pp. 8176–8185.

[3] 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)*, June 2024, pp. 7017–7026.

[4] 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,” *Advances in neural information processing systems*, vol. 30, 2017.

[5] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[6] Z. WANG, “Image quality assessment : Form error visibility to structural similarity,” *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 604–606, 2004.

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

[8] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy, “Exploring clip for assessing the look and feel of images,” in *Proceedings of the AAAI conference on artificial intelligence*, 2023, vol. 37, pp. 2555–2563.

[9] Sijing Wu, Yunhao Li, Ziwen Xu, Yixuan Gao, Huiyu Duan, Wei Sun, and Guangtao Zhai, “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*, 2025, pp. 6928–6937.

[10] Yixuan Gao, Xiongkuo Min, Jinliang Han, Yaqin Cao, Sijing Wu, Yunze Dou, and Guangtao Zhai, “Multi-dimensional text-to-face image quality assessment using lmm: Database and method,” in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 6948–6957.

[11] Yunhao Li, Sijing Wu, Wei Sun, Zhichao Zhang, Yucheng Zhu, Zicheng Zhang, Huiyu Duan, Xiongkuo Min, and Guangtao Zhai, “Aghi-qa: A subjective-aligned dataset and metric for ai-generated human images,” *arXiv preprint arXiv:2504.21308*, 2025.

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

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

[14] Yingjie Zhou, Zicheng Zhang, Sijing Wu, Jun Jia, Yanwei Jiang, Wei Sun, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai, “Mi3s: A multimodal large language model assisted quality assessment framework for ai-generated talking heads,” *Information Processing & Management*, vol. 63, no. 1, pp. 104321, 2026.

[15] Huiyu Duan, Kang Fu, Sijing Wu, Yunhao Li, Zicheng Zhang, Qiang Hu, Xiongkuo Min, and Guangtao Zhai, “Bmpcq: Bioinspired metaverse point cloud quality assessment based on large multimodal models,” *Advanced Intelligent Systems*, p. 2500504, 2025.

[16] Sijing Wu, Yunhao Li, Huiyu Duan, Yanwei Jiang, Yucheng Zhu, and Guangtao Zhai, “Hveval: Towards unified evaluation of human-centric video generation and understanding,” in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 13376–13383.

[17] Xianbing Sun, Yan Hong, Jiahui Zhan, Jun Lan, Huijia Zhu, Weiqiang Wang, Liqing Zhang, and Jianfu Zhang, “Ds-vton: High-quality virtual try-on via disentangled dual-scale generation,” *arXiv preprint arXiv:2506.00908*, 2025.

[18] Sen He, Yi-Zhe Song, and Tao Xiang, “Style-based global appearance flow for virtual try-on,” in *CVPR*, 2022.

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

[20] Jianhao Zeng, Dan Song, Weizhi Nie, Hongshuo Tian, Tongtong Wang, and An-An Liu, “Cat-dm: Controllable accelerated virtual try-on with diffusion model,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8372–8382.

[21] Yuhao Xu, Tao Gu, Weifeng Chen, and Chengcai Chen, “Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on,” *arXiv preprint arXiv:2403.01779*, 2024.

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

[23] Kuashou Technology, “Kling ai virtual try-on,” 2024.

[24] LinkFox, “Linkfox ai dressing: Virtual try-on platform,” 2024.

[25] Mostafa Atef, Mariam Ayman, Ahmed Rashed, Ashrakat Saeed, Abdelaahman Saeed, and Ahmed Fares, “Efficientviton: An efficient virtual try-on model using optimized diffusion process,” *arXiv preprint arXiv:2501.11776*, 2025.

[26] Haoyu Wang, Zhiyu Zhang, Donglin Di, Shiliang Zhang, and Wangmeng Zuo, “Mv-vton: Multi-view virtual try-on with diffusion models,” *arXiv preprint arXiv:2404.17364*, 2024.

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

[28] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton, “Demystifying mmd gans,” *arXiv preprint arXiv:1801.01401*, 2018.

[29] RIR BT, “Methodology for the subjective assessment of the quality of television pictures,” *International Telecommunication Union*, vol. 4, pp. 19, 2002.

[30] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang, “Fsim: A feature similarity index for image quality assessment,” *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[31] Ke Gu, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang, “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*, 2012, pp. 1–6.

[32] Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C. Bovik, “Gradient magnitude similarity deviation: A highly efficient perceptual image quality index,” *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 684–695, 2014.

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

[34] Shanshan Lao, Yuan Gong, Shuwei Shi, Sidi Yang, Tianhe Wu, Jiahao Wang, Weihao Xia, and Yujiu Yang, “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*, 2022, pp. 1140–1149.

[35] Le Kang, Peng Ye, Yi Li, and David Doermann, “Convolutional neural networks for no-reference image quality assessment,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[36] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek, “Deep neural networks for no-reference and full-reference image quality assessment,” *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 206–219, 2018.

[37] Hossein Talebi and Peyman Milanfar, “Nima: Neural image assessment,” *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3998–4011, 2018.

[38] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang, “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*, 2020, pp. 3667–3676.

[39] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin, “Topiq: A top-down approach

from semantics to distortions for image quality assessment,” *IEEE Transactions on Image Processing*, vol. 33, pp. 2404–2418, 2024.