
VTON-VLLM: Aligning Virtual Try-On Models with Human Preferences

—NeurIPS 2025 Supplementary Material

Anonymous Author(s)

Affiliation

Address

email

1 A More Quantitative and Qualitative Results

2 **Quantitative Results on DressCode.** The quantitative results on DressCode across three macro-
3 categories are summarized in Table 1 and Table 2. Table 1 reports SSIM, LPIPS, GC_p , and IQ_p under
4 the *paired* setting, while Table 2 presents results under the *unpaired* setting, including FID, KID,
5 GC_u , and IQ_u . With the integration of our proposed VRM-Instruct and TTS modules, state-of-the-art
6 VTON models achieve further improvements on the DressCode benchmark. Specifically, in the paired
7 setting for the upper-body category, **OOTDiffusion w/ VRM-Instruct + TTS** increases SSIM from
8 0.902 to 0.928 and GC_p from 3.59 to 4.14, indicating enhanced detail preservation. Meanwhile, in
9 the unpaired setting, the same recipe boosts IQ_u from 4.05 to 4.15 and reduces FID from 11.03 to
10 9.43, which demonstrates improved visual realism. These results on the DressCode dataset further
11 validate the effectiveness of our VRM-Instruct and TTS modules in guiding VTON models towards
12 better alignment with human preferences in both image quality and garment fidelity.

13 **More Qualitative Results.** In Figure 1, we present additional qualitative comparisons of synthesized
14 results produced by different methods on VITON-HD. Furthermore, more results for other clothing
15 categories (i.e., lower-body garments and dresses) on DressCode are presented in Figure 2.

16 B More Discussions

17 **Effect of Iteration Count N in Test-Time Scaling.** We analyze the sensitivity of our proposed TTS
18 with respect to the iteration count N , which is guided by human-preference-aligned reward from
19 VTON-VLLM. The results are reported in Table 3, and the best results are achieved with $N = 3$.

20 **VTON-VLLM vs Base VLLM.** To further evaluate the effectiveness of our proposed VTON-VLLM,
21 we compare its capability in VTON assessment with that of the base VLLM (i.e., Pixtral-12B [1])
22 without fine-tuning on human feedback data. We sample 100 challenging garment-person pairs
23 from VITON-HD, where OOTDiffusion [9] struggles to produce satisfactory outputs, and generates
24 10 candidate images for each pair using different random seeds. Human annotators then compare
25 the highest-rated images selected by both models from the 10 candidates to determine which is
26 better. As a result, our VTON-VLLM achieves a win rate of 74.17% compared to the base VLLM.
27 Some examples are shown in Figure 4. It is evident that our VTON-VLLM consistently selects the
28 perceptually superior image, whereas the base VLLM often overlooks subtle garment inconsistency
29 and visual artifacts. This highlights the effectiveness of our fine-tuned VTON-VLLM in aligning
30 VTON evaluations with human judgments.

Table 1: Quantitative results on DressCode in paired setting of three categories(D.C.Upper, D.C.Lower, D.C.Dress). VRM-Instruct and TTS are short for VTON refinement model with fine-grained instructions and test-time scaling, respectively.

Train/Test Method	D.C.Upper/D.C.Upper				D.C.Lower/D.C.Lower				D.C.Dress/D.C.Dress			
	SSIM \uparrow	LPIPS \downarrow	GC _p \uparrow	IQ _p \uparrow	SSIM \uparrow	LPIPS \downarrow	GC _p \uparrow	IQ _p \uparrow	SSIM \uparrow	LPIPS \downarrow	GC _p \uparrow	IQ _p \uparrow
LaDI-VTON [5]	0.904	0.063	3.48	3.32	0.862	0.122	3.36	3.22	0.782	0.129	3.25	3.02
w/ VRM-Instruct and TTS	0.930	0.042	4.12	3.99	0.923	0.043	4.06	3.81	0.912	0.069	3.97	3.76
OOTDiffusion [9]	0.902	0.050	3.59	3.43	0.882	0.103	3.42	3.26	0.824	0.100	3.40	3.15
w/ VRM-Instruct and TTS	0.928	0.039	4.14	3.92	0.918	0.044	4.01	3.76	0.898	0.064	3.92	3.74
IDM-VTON [3]	0.911	0.060	3.70	3.57	0.913	0.055	3.60	3.33	0.863	0.082	3.62	3.44
w/ VRM-Instruct and TTS	0.931	0.034	4.08	3.98	0.917	0.040	3.92	3.72	0.883	0.062	3.94	3.72
SPM-Diff [8]	0.927	0.042	3.74	3.60	0.914	0.050	3.70	3.45	0.892	0.073	3.68	3.49
w/ VRM-Instruct and TTS	0.928	0.035	4.06	4.03	0.920	0.046	4.03	3.79	0.890	0.069	3.91	3.70
CAT-VTON [4]	0.919	0.045	3.68	3.58	0.897	0.063	3.65	3.40	0.900	0.065	3.54	3.41
w/ VRM-Instruct and TTS	0.927	0.039	4.11	4.00	0.918	0.042	3.99	3.74	0.892	0.065	3.93	3.76

Table 2: Quantitative results on DressCode in unpaired setting of three categories(D.C.Upper, D.C.Lower, D.C.Dress). VRM-Instruct and TTS are short for VTON refinement model with fine-grained instructions and test-time scaling, respectively.

Train/Test Method	D.C.Upper/D.C.Upper				D.C.Lower/D.C.Lower				D.C.Dress/D.C.Dress			
	FID \downarrow	KID \downarrow	GC _u \uparrow	IQ _u \uparrow	FID \downarrow	KID \downarrow	GC _u \uparrow	IQ _u \uparrow	FID \downarrow	KID \downarrow	GC _u \uparrow	IQ _u \uparrow
LaDI-VTON [5]	14.26	3.33	4.02	3.78	13.38	1.98	2.86	2.64	13.12	1.85	3.21	2.99
w/ VRM-Instruct and TTS	9.38	0.52	4.34	4.19	9.80	0.85	3.27	3.10	10.65	0.84	3.40	3.55
OOTDiffusion [9]	11.03	0.86	4.10	4.05	9.62	0.84	2.99	2.92	10.65	0.84	3.40	3.29
w/ VRM-Instruct and TTS	9.43	0.42	4.31	4.15	9.66	0.80	3.19	3.08	9.69	0.52	3.76	3.58
IDM-VTON [3]	10.86	0.62	4.13	4.11	12.05	0.93	3.08	3.02	12.33	1.41	3.48	3.29
w/ VRM-Instruct and TTS	9.55	0.43	4.30	4.18	9.74	0.77	3.28	3.03	9.92	0.54	3.80	3.51
SPM-Diff [8]	10.56	0.40	4.23	4.15	9.02	0.80	3.11	3.02	10.17	0.50	3.55	3.49
w/ VRM-Instruct and TTS	9.46	0.45	4.32	4.20	9.18	0.78	3.26	3.11	9.79	0.51	3.78	3.54
CAT-VTON [4]	8.92	0.51	4.16	4.12	9.21	0.94	3.06	2.98	9.76	0.66	3.50	3.41
w/ VRM-Instruct and TTS	9.23	0.44	4.35	4.14	9.20	0.76	3.19	3.12	9.68	0.57	3.83	3.61

Table 3: Ablation study of iteration count N in our TTS on VITON-HD.

Model	Garment Consistency				Image Quality			
	Pattern	Characters	Sleeve	Shape	Edge Artifact	Human Pose	SSIM \uparrow	LPIPS \downarrow
2	4.817	4.803	4.720	4.688	4.736	4.450	0.825	0.064
3	4.931	4.912	4.822	4.872	4.890	4.726	0.903	0.043
5	4.896	4.882	4.807	4.750	4.820	4.601	0.883	0.058

31 C Dataset Details

32 **Human Feedback Data.** During the collection of human feedback data, a pre-trained VLLM is
33 first employed to roughly assess garment-person images across various fine-grained dimensions.
34 Human annotators then review and validate these evaluations to ensure accuracy. These annotations
35 are further structured into multi-turn conversational instructions for training our VTON-VLLM. An
36 example instruction for text character consistency is illustrated in Figure 3.

37 **Synthetic VITON-SYN.** We adopt a similar network to our VRM-Instruct (described in Section 4.1)
38 as the VTON model and train it on VITON-HD to generate VTON data in Section 5.3. Different
39 from VRM-Instruct, the inpainting condition in this VTON model is obtained by combining the
40 garment image I_g and a blank white image I_1 , leading to $I_{cond} = I_g \odot I_1$. To obtain VITON-SYN,
41 we sample garment images from existing datasets [2, 6, 7] and scene prompts from a pre-defined
42 pool as reference conditions, and synthesize images of persons wearing the corresponding garments
43 with the trained VTON model. The derived garment-person pairs are further rigorously filtered
44 by the proposed VTON-VLLM, resulting in a high-quality and human-preference-aligned training
45 dataset. Figure 5 illustrates the data construction pipeline. Overall, VITON-SYN consists of 100,000
46 garment-person pairs spanning ten different scenarios, which is significantly larger and more diverse
47 than VITON-HD dataset (13,679 indoor samples). Figure 6 showcases various prompts along with
48 their corresponding synthesized images from VITON-SYN.

49 **New Test Set VITON-Bench.** Unlike existing benchmarks such as VITON-HD and DressCode,
50 our VITON-Bench emphasizes more on challenging VTON scenarios. These include uniquely
51 styled clothing, complex human poses (e.g., sitting or lying down), and diverse, intricate outdoor
52 environments. Figure 7 illustrates some examples from our VITON-Bench, highlighting its greater
53 scene diversity and increased difficulty compared to previous benchmarks (see Figure 1 and 2).



Figure 1: Qualitative results on VTON-HD dataset in unpaired settings. For each pair, the left image is generated by a baseline model (i.e., LaDI-VTON, IDM-VTON, CAT-VTON), and the right image is enhanced by incorporating our proposed VRM-Instruct and TTS module powered by VTON-VLLM.



Figure 2: Qualitative results on DressCode dataset in unpaired settings. For each pair, the left image is generated by a baseline model (i.e., LaDI-VTON, IDM-VTON, CAT-VTON), and the right image is enhanced by incorporating our proposed VRM-Instruct and TTS module powered by VTON-VLLM.

System-Message(X_0): You are a professional digital artist. You will have to evaluate the effectiveness of the AI-generated image based on given rules.

Human(X_q^1): Please describe characters on upper-body garment in each of two images separately.

Assistant(X_a^1): In the first image, the upper-body garment is a horizontally striped red and white T-shirt. The word "FIORUCCI" is printed in bold, black and red letters across the chest area, centered and slightly tilted. In the second image, the T-shirt is the same as in the first image, featuring the horizontally striped red and white pattern. The word "FIORUCCI" is also printed in bold, black and red letters across the chest area, centered and slightly tilted.

Human(X_q^2): Now, based on the characters on the upper-body garments in the two images, please determine how well the characters are consistent in the two images, output a score in [0, 5].

Assistant(X_a^2): 5.

Human(X_q^3): Review your previous responses step-by-step. If you are completely confident that your reasoning and conclusions are correct, respond with "Correct". If you find any errors or are uncertain, respond with "Wrong".

Assistant(X_a^3): Correct.

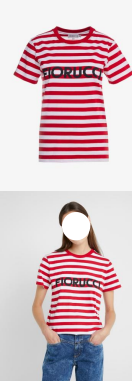


Figure 3: An example of a multi-turn conversational instruction for text character consistency from our human feedback dataset.



Figure 4: Comparisons between top-ranked images selected by our VTON-VLLM and the base VLLM from 10 candidate synthesized images for each garment-human pair.

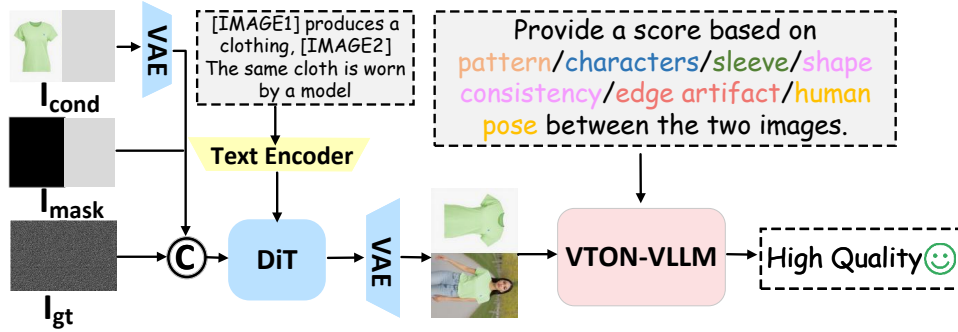


Figure 5: Data construction pipeline of synthetic VITON-SYN. The instruction provided to VTON-VLLM is simplified here.

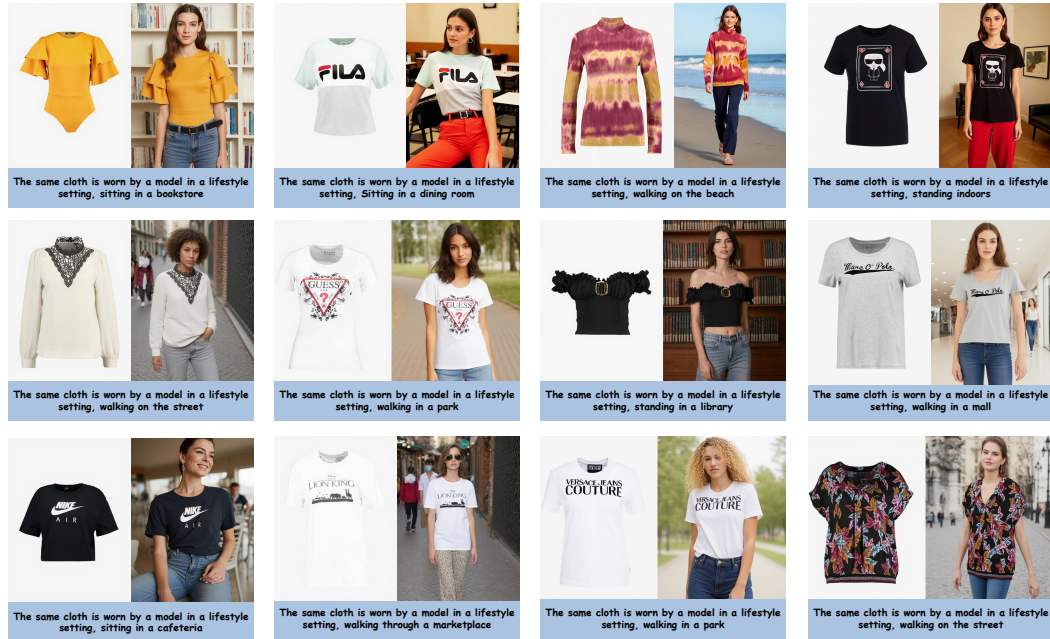


Figure 6: Different prompts along with their corresponding synthesized images from VITON-SYN.

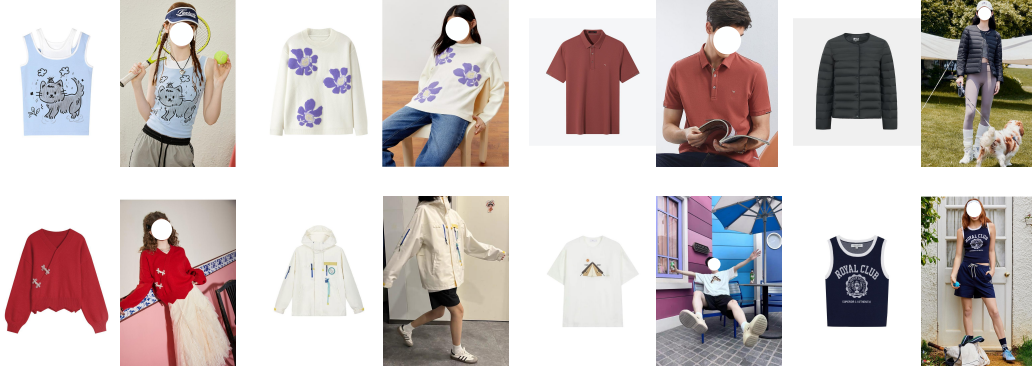


Figure 7: Examples from our VITON-Bench.

References

- [1] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024.
- [2] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *CVPR*, 2021.
- [3] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for authentic virtual try-on in the wild. In *ECCV*, 2024.
- [4] Zheng Chong, Xiao Dong, Haoxiang Li, shiyue Zhang, Wenqing Zhang, Hanqing Zhao, xujie zhang, Dongmei Jiang, and Xiaodan Liang. CatVTON: Concatenation is all you need for virtual try-on with diffusion models. In *ICLR*, 2025.
- [5] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. In *ACM MM*, 2023.
- [6] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: High-resolution multi-category virtual try-on. In *CVPR*, 2022.
- [7] Fei Shen, Xin Jiang, Xin He, Hu Ye, Cong Wang, Xiaoyu Du, Zechao Li, and Jinhui Tang. Imagdressing-v1: Customizable virtual dressing. *arXiv preprint arXiv:2407.12705*, 2024.
- [8] Siqi Wan, Jingwen Chen, Yingwei Pan, Ting Yao, and Tao Mei. Incorporating visual correspondence into diffusion model for virtual try-on. In *ICLR*, 2025.
- [9] 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.