1. Supplementary

1.1. TryOn-Refiner (SD3 version)

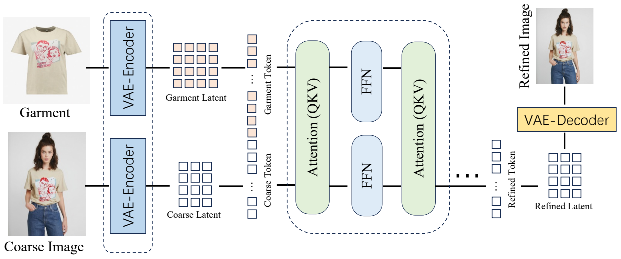

Figure 1. The pipeline of our SD3-based TryOn-Refiner.

Architecture: In the main part of our paper, we introduce the SDXL-based TryOn-Refiner. To enable a comprehensive analysis of the performance of the conditional rectified-flow-based TryOn Refiner within a transformer architecture, we have also developed an SD3-based TryOn-Refiner, as illustrated in Fig. 1. Starting with a coarse image $X_1 \in R^{1024\times768}$ and a garment image $C_g R^{1024\times1024}$, we first encode these into the latent feature $L_{coarse} \in R^{128\times96\times16}$and $L_g \in R^{128\times128\times16}$ using the VAE encoder of SD3. These latent features are then processed with a 2×2 patching operation and reshaped into the coarse token $T_{coarse} \in R^{m\times16}$ and the garment token $T_{garment} \in R^{n\times16}$.

It is noteworthy that our conditional rectified-flow-based TryOn Refiner is initialized from a coarse image; thus, in the SD3 version of the network, the original noise token has been removed. Moreover, because our TryOn Refiner places greater emphasis on visual information and is less sensitive to textual information, we have eliminated the original text token. So the final input of the attention process is the combination of the coarse token T_{coarse} and the garment token $T_{garment}$ as shown in Fig. 1:

It is noteworthy that our conditional rectified-flow-based TryOn Refiner is initialized with a coarse image. Consequently, in the SD3 version of the network, the original noise token has been removed. Furthermore, as our TryOn Refiner places greater emphasis on visual information and is less sensitive to textual data, the original text token has also been excluded. Therefore, the final input to the attention process is a combination of the coarse token T_{coarse} and the garment token $T_{garment}$, as illustrated in Fig. 1:

$$\text{Attention} = \text{softmax}\left(\frac{(Q_c, Q_g)\cdot \text{cat}(K_c, K_g)^T}{\sqrt{d_k}}\right)\cdot \text{cat}(V_c, V_g), \tag{1}$$

where c and g denote the $Q/K/V$ of the coarse image and the garment image, respectively. After a series of computations, the network ultimately outputs the refined coarse token, which we refer to as the refined token. The SD3-based TryOn-Refiner then performs an inverse patch operation on the refined token to obtain the refined latent representation. Finally, this refined latent representation is passed through the VAE-Decoder to generate the final refined image.

Method	LPIPS ↓	SSIM ↑	FID ↓	KID ↓
VITON-HD [1]	0.116	0.863	12.13	3.22
HR-VITON [5]	0.097	0.878	12.30	3.82
LaDI-VTON [6]	0.091	0.875	9.31	1.53
GP-VTON [7]	0.083	0.892	9.17	0.93
TryOn-Adapter [8]	0.071	0.894	**8.63**	**0.79**
StableVITON [4]	0.084	0.862	9.13	1.20
OOTD [9]	0.071	0.878	8.81	0.82
SDXL-TryOn	0.067	0.885	9.07	0.95
+TryOn-Refiner	**0.061**	**0.895**	8.70	0.83

Table 1. Quantitative comparisons of our SD3-based TryOn-Refiner on the VITON-HD benchmark, where the best and second-best results are reported in bold and underlined, respectively.

Comparison with other methods: Table 1 presents the quantitative evaluation results of our SD3-based TryOn-Refiner on the VITON-HD dataset. Firstly, we compare our method (SDXL-TryOn) quantitatively with previous traditional methods on these datasets, including VITON-HD [1], HR-VITON [5], LaDI-VTON [6], GP-VTON [7], TryOn-Adapter [8], OOTD [9], and StableVITON [4].

Based on the experimental results presented in Table 1, we draw the following conclusions: 1) When comparing our SDXL-TryOn with existing try-on methods listed in the table, it does not show significant advantages. This is because SDXL-TryOn acts primarily as a foundational method with a robust baseline (SDXL). 2) Even when built upon a relatively lower-performing try-on model, our SD3-based TryOn-Refiner enhances the refined results to a level nearly reaching state-of-the-art (SOTA) standards. For instance, the LPIPS score of 0.061 surpasses all competitors, the SSIM score of 0.895 also outperforms all competitors, and the FID score of 8.70 ranks second only to TryOn-Adapter's 8.63, despite TryOn-Adapter employing several additional constraints.

Visualiation: To better demonstrate the detail enhancement capabilities of our SD3-based TryOn-Refiner, we have included a substantial number of visualizations in this section.

As shown in Fig. 5, our SD3-based TryOn-Refiner demonstrates exceptional capabilities in restoring texture details. For example, the top worn by the model in the first row features complex textures around the neckline, which traditional try-on methods fail to represent accurately (Coarse Detail). Notably, our SD3-based TryOn-Refiner successfully restores these details, enhancing customer satisfaction during the virtual try-on process.

Moreover, Fig. 6 presents visualizations of the text detail

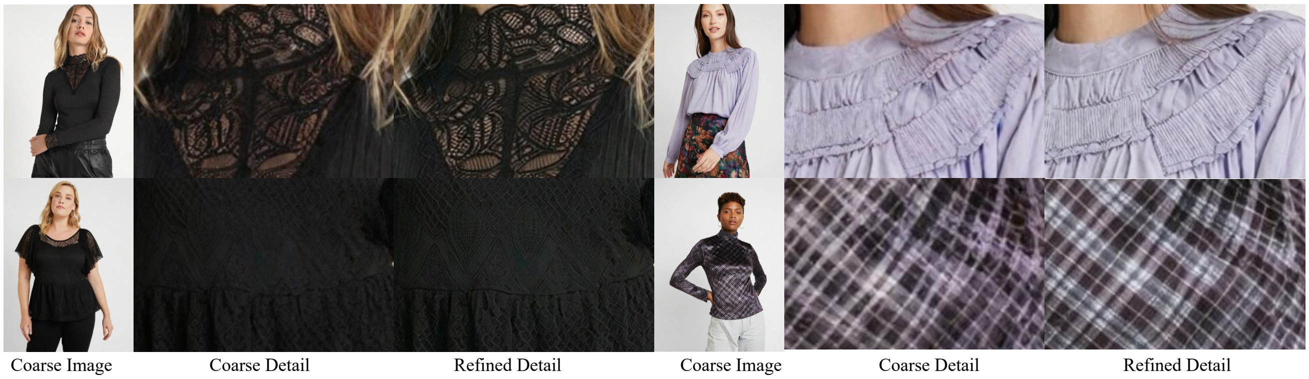

Figure 2. Texture Detail Restoration Visualization of our SD3-based TryOn-Refiner.

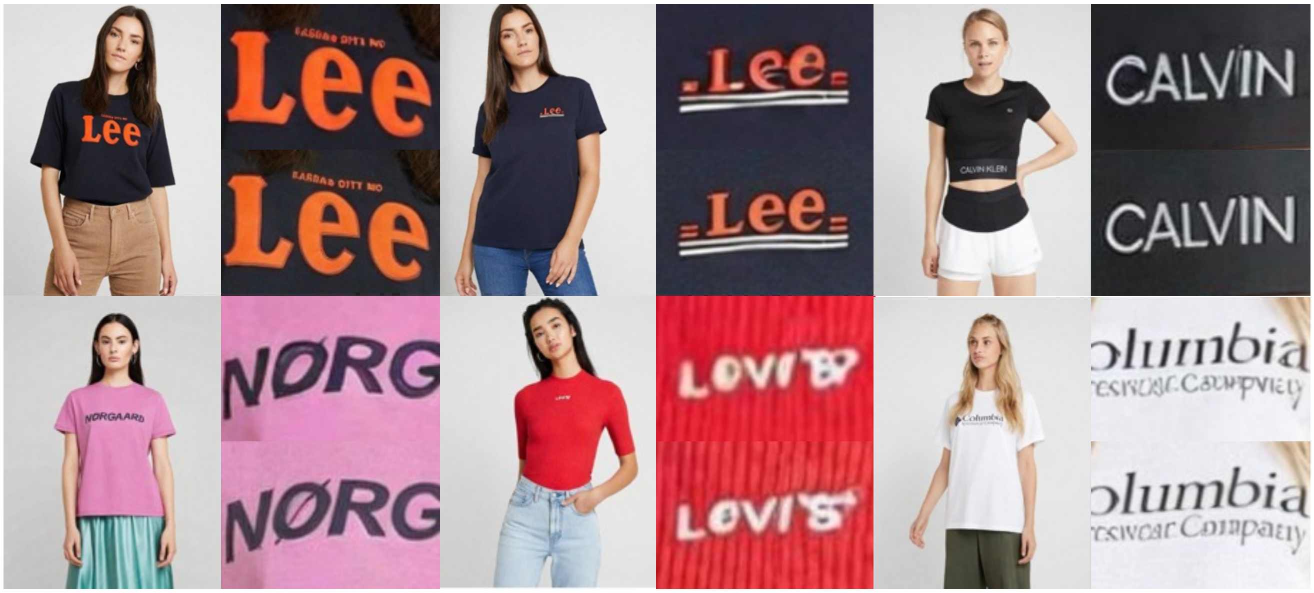

Figure 3. Text Detail Restoration Visualization of our SD3-based TryOn-Refiner.

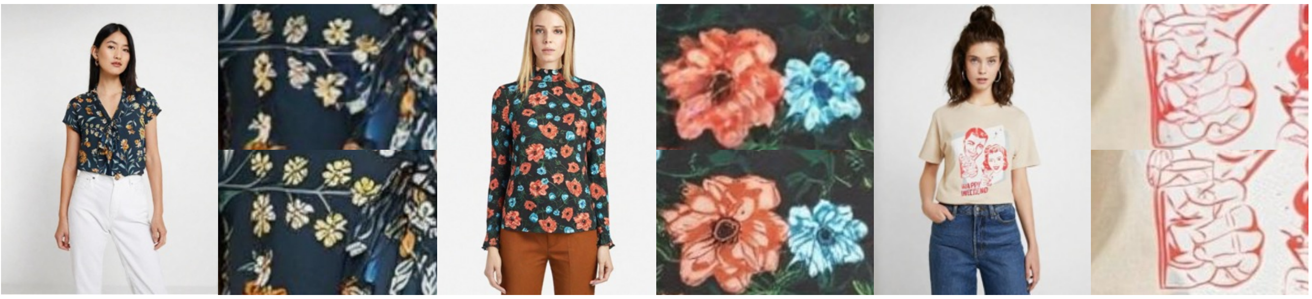

Figure 4. Pattern Detail Restoration Visualization of our SD3-based TryOn-Refiner.

restoration capabilities of the SD3-based TryOn-Refiner. Each subfigure comprises three parts: the left side displays the try-on results, the top right highlights the coarse details from the first-stage try-on method, and the bottom right shows the refined details from our TryOn-Refiner. We draw the following conclusions: 1) The SD3-based TryOn-Refiner demonstrates strong restoration capabilities for text, effectively restoring even small text, such as the letters ”MO” in the first image. 2) However, due to the complexity of the text, we note that in some cases—especially when

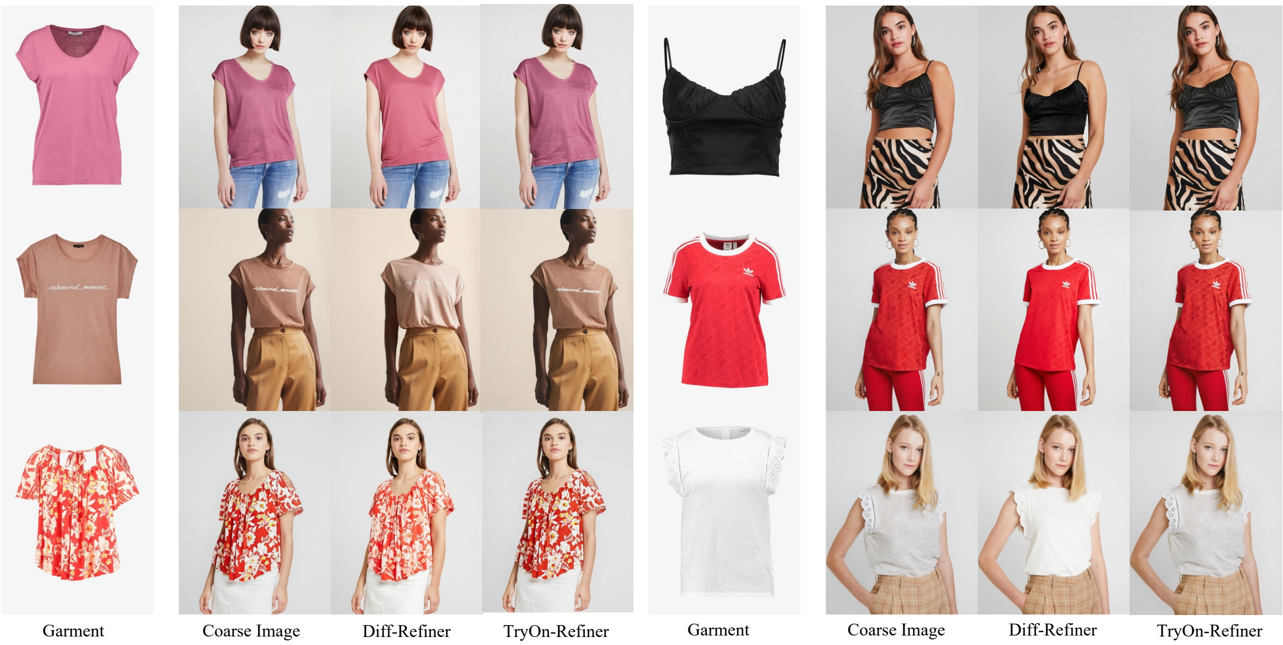

Figure 5. The qualitative comparison between Diff-Refiner and TryOn-Refiner (Color deviation).

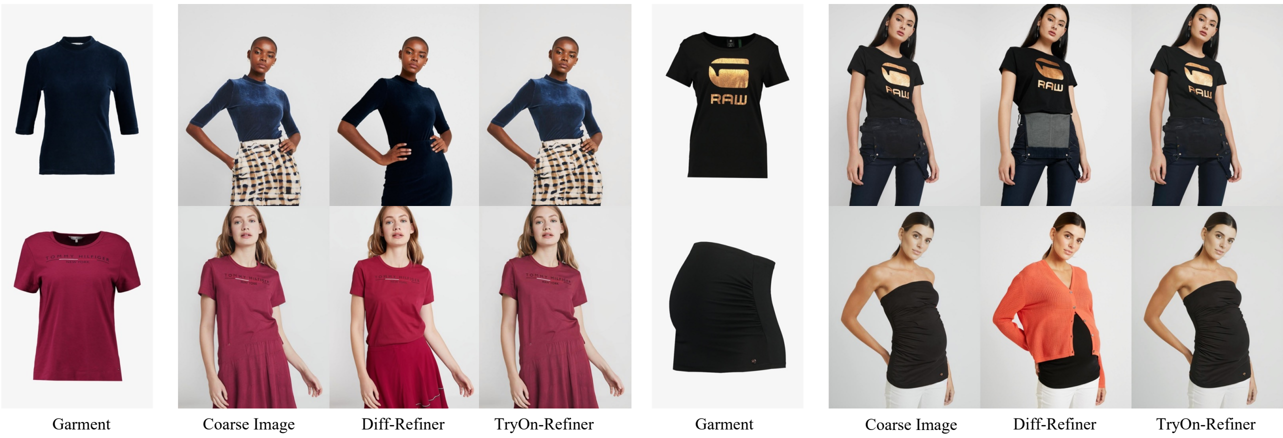

Figure 6. The qualitative comparison between Diff-Refiner and TryOn-Refiner (Background difference).

the text is very small or the reference image's text is unclear—the TryOn-Refiner may not achieve perfect restoration. Nevertheless, in these instances, the refined results are still significantly clearer and more recognizable than the initial try-on results, as shown by the last two examples in the second row.

Finally, Fig. 7 illustrates the pattern detail restoration capabilities of the SD3-based TryOn-Refiner. For instance, the floral patterns in the first and second columns have been beautifully restored. Additionally, in the third example, the hand of the cartoon character holding the beer mug has been exceptionally well restored.

1.2. More Qualitative Visualization between Diff-Refiner and TryOn-Refiner.

As discussed in our main paper, the diffusion-based refinement model can restore details similarly to our TryOn-Refiner. However, the diffusion-based model has several drawbacks:

1) The diffusion-based refinement model is prone to color discrepancies and the loss of certain original details. As illustrated in Fig. 5, the refined images in row 1 exhibit noticeable color inconsistencies compared to the original images, while the image in row 2, column 2, demonstrates a degradation of details.

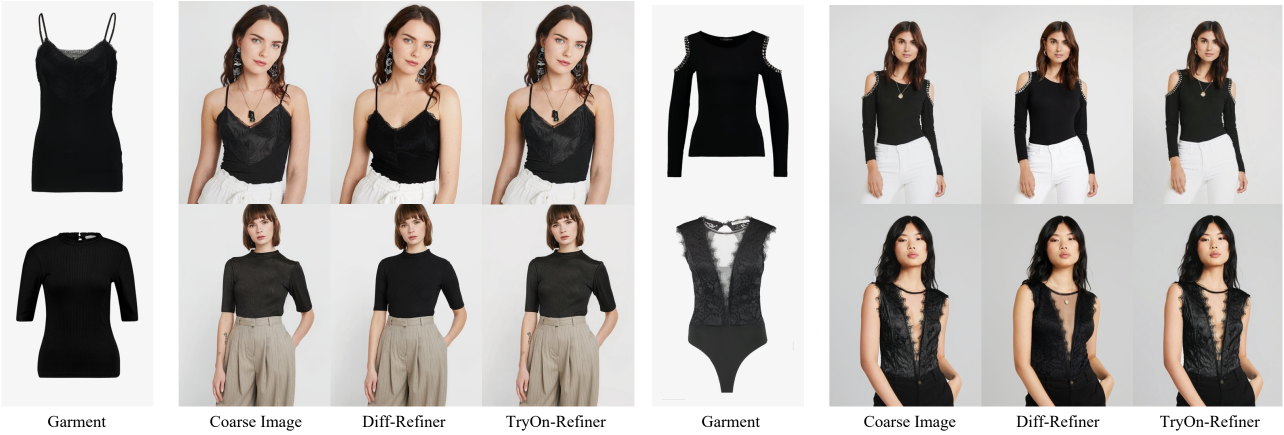

Figure 7. The qualitative comparison between Diff-Refiner and TryOn-Refiner (Detail difference).

2) Due to the uncertainty introduced by starting from noise during the generation process, the diffusion-based refinement model often results in changes to areas outside the intended regions. Fig. 6 presents several such cases: a) In the first image, the diffusion-based refinement model mistakenly changes the model's pants to match the style of the top, a significant error that does not occur with our TryOn-Refiner. b) In the second image, the model incorrectly introduces some background information. c) In the third image, after using the diffusion-based refinement model, the red dress the model is wearing exhibits white streaks. d) In the last image, the model is mistakenly dressed in an orange coat.

3) In addition to the previously mentioned issues, we observed that the diffusion-based refinement model sometimes alters details, such as the model's accessories, which may lead to user dissatisfaction. Fig. 7 highlights some of these instances: a) In the first, second, and fourth images, the model's necklace is missing; b) In the third image, the tattoo on the model's right hand has disappeared.

However, as illustrated in Fig. 5, Fig. 6, and Fig. 7, our TryOn-Refiner does not exhibit issues such as color discrepancies, background changes, or alterations in details. This demonstrates that the conditional rectified-flow-based TryOn-Refiner is more stable and reliable than the diffusion-based refinement model.

1.3. More Comparison.

We apply TryOn-Refiner to CAT-VTON (FLUX) [3] and IDM-VTON [2] for validating the effectiveness of our method. We engaged around 5 reviewers to conduct user studies on the test set (2032 images) of VITON-HD benchmark. The reviewers were asked to vote on whether the refined result is "good", "same", or "bad" compared to the coarse try-on result. For these two different try-on methods, our TryOn-Refiner achieved improvements in 16.4% 17.8% of the cases, while less than 1% of the cases worsened.

Method	Good	Similarity	Bad
CAT-VTON (FLUX)	16.4%	82.9%	0.7%
IDM-VTON	17.8%	82.2%	0%

Table 2. A user study of Our TryOn-Refiner Integrated with Other Methods.

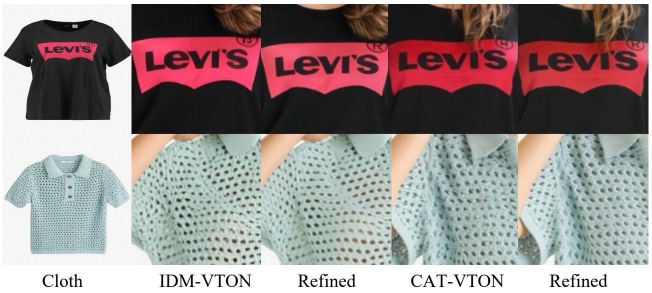

Figure 8. Comparative Visualization of Our TryOn-Refiner Integrated with Other Methods.

References

[1] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14131–14140, 2021. 1

[2] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for authentic virtual try-on in the wild. In *European Conference on Computer Vision*, pages 206–235. Springer, 2024. 4

[3] Zheng Chong, Xiao Dong, Haoxiang Li, Shiyue Zhang, Wenqing Zhang, Xujie Zhang, Hanqing Zhao, Dongmei Jiang, and Xiaodan Liang. Catvton: Concatenation is all you need

for virtual try-on with diffusion models. *arXiv preprint arXiv:2407.15886*, 2024. 4

[4] 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*, pages 8176–8185, 2024. 1

[5] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *European Conference on Computer Vision*, pages 204–219. Springer, 2022. 1

[6] 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 31st ACM International Conference on Multimedia*, pages 8580–8589, 2023. 1

[7] Zhenyu Xie, Zaiyu Huang, Xin Dong, Fuwei Zhao, Haoye Dong, Xijin Zhang, Feida Zhu, and Xiaodan Liang. Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23550–23559, 2023. 1

[8] Jiazheng Xing, Chao Xu, Yijie Qian, Yang Liu, Guang Dai, Baigui Sun, Yong Liu, and Jingdong Wang. Tryon-adapter: Efficient fine-grained clothing identity adaptation for high-fidelity virtual try-on. *arXiv preprint arXiv:2404.00878*, 2024. 1

[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. 1