M&M VTO: Multi-Garment Virtual Try-On and Editing

Supplementary Material

1. Implementation Details

1.1. Training and inference

M&M VTO is trained in two stages. For the first stage, the model is trained on 512×256 images for $600K$ iterations. In the second stage, the model is initialized from the pretrained checkpoint of the first stage and trained on 1024×512 images for an additional $200K$ iterations. For both training stages, the batch size is set to 1024, and the learning rate linearly increases from 0 to 10^{-4} in the first $10K$ steps and is kept unchanged afterwards. We parameterize the model output in v-space following [13] while the $L2$ loss is computed in ϵ-space. All conditional inputs are set to 0 in 10% of the training time for classifier-free guidance (CFG) [5]. Test results are generated by sampling M&M VTO for 256 steps using ancestral sampler [6].

1.2. Garment attributes

We summarize as follows the full set of attributes used as layout conditioning input y_{gl}.

1. What is the type of the sleeve?
   (a) Not applicable
   (b) Sleeveless
   (c) Short sleeve
   (d) Middle sleeve
   (e) Long sleeve
2. Is the sleeve rolled up?
   (a) Not applicable
   (b) Sleeve type is not long
   (c) Yes
   (d) No
3. Is the top garment tucked in?
   (a) Not applicable
   (b) Not wearing top garment
   (c) Can not determine
   (d) Yes
   (e) No
4. Is the person wearing outer top?
   (a) Not applicable
   (b) Yes
   (c) No
5. Is the outer top closed (*e.g.* zipper up or button on)?
   (a) Not applicable
   (b) Not wearing outer top
   (c) Can not determine
   (d) Yes
   (e) No

We selected $1,500$ images and asked human labelers to answer all questions for each image. After that, we converted question-answer pairs into a formatted text, where different question-answer pairs are separated by semicolon while the question and answer within each pair are separated by colon. The resulting $1,500$ image-caption samples were used to finetune PaLI-3 [2] model. Finally, we ran inference of the finetuned model on our train and test data, and converted the formatted text back into class labels.

Methods	FID ↓	KID ↓
GP-VTON [15]	38.392	33.909
LaDI-VTON [11]	19.346	9.305
Ours-DressCode	**18.725**	**8.250**

Table 1. Our method trained solely on DressCode vs GP-VTON and LaDI-VTON official checkpoints. We report FID and KID on DressCode triplets test set.

2. Results

In this section, we provide additional qualitative and quantitative results.

2.1. Comparison of VTO

In Figure 1, 2, 3 and 4, we showcase additional qualitative results from our $8,300$ triplets test set, comparing them against those generated by TryOnDiffusion [16], where both methods are trained on our "garment paired" and "layflat paired" dataset. These results highlight our method's superior ability to retain garment details and layout. We also compare to "layflat-VTO" methods GP-VTON [15] and LaDI-VTON [11] on DressCode [10] triplets test dataset. To ensure a fair comparison, we trained our method exclusively on the DressCode dataset. The FID and KID metrics for the DressCode triplets test set, presented in Table 1, demonstrate that our method surpasses GP-VTON and LaDI-VTON in both metrics, even when **trained solely** on the DressCode dataset. Further qualitative comparisons on the DressCode triplets test set against all baselines are provided in Figure 5 and 6.

2.2. Comparison of Editing

We conducted a user study with 200 images to compare garment layout editing. The results in Table 2 indicate that our method are preferred by users 84.5% of the time, outperforming the baseline methods. Figure 7, 8, 9 and 10 present qualitative comparisons on different layout editing tasks. These examples demonstrate our method's ability to perform the intended edits accurately while preserving the integrity of other areas in both the person and the garments.

Image editing baselines require different sets of inputs, such as masks. InstructPix2Pix [1] and Prompt-to-Prompt

Methods	US ↑
P2P + NI [9]	0
IP2P [1]	1
Imagen editor [14]	10
DiffEdit [3]	0
SDXL inpainting [12]	4
Ours	**169**
Hard to tell	16

Table 2. **User Study for try-on editing.** We conducted user study on 200 images. The users are required to select the best method that can successfully perform the editing task while maintaining the property of input person and garments.

Methods	US ↑
Finetuned full model	19
Finetuned person encoder	20
Ours without finetuning	95
Ours with finetuning	**265**
Hard to tell	1

Table 3. **User Study for person finetuning.** We carried out a user study involving 400 images across 4 subjects, where we randomly select 100 top + bottom input garments for each subject. The participants were asked to choose the method that best maintains the identity of the person (including body pose and shape) as well as the details of input garments.

Methods	FID ↓	KID ↓
Cascaded	18.523	**15.218**
From Scratch	21.645	15.781
Ours	**18.145**	15.227

Table 4. **Quantitative results for ablation studies.** We report FID and KID on our $8,300$ triplets test set.

(P2P) [4] with null inversion [9] only requires text editing instructions. DiffEdit [3], Imagen Editor [14], and Stable Diffusion XL Inpainting [12] require masks for the region of interest. To automatically obtain masks for image editing, we use human pose estimations to mask out belly regions for "tuck in top garment" or "tuck out top garment" or the arm regions for "roll up sleeve" or "roll down sleeve".

2.3. Finetuning Comparison

We chose 4 person images with challenging body shapes or poses for our person finetuning comparison. For each person image, we randomly picked 100 top and bottom garment combinations, then generated try-on results using all baseline methods as well as our own. The user study results, detailed in Table 3, show our finetuning method significantly outperforming the baselines. Additionally, Figure 11, 12, 13 and 14 showcase qualitative comparison for each subject. Without finetuning, the person's arms, legs, or torso may appear unnaturally slim or wide, and certain challenging poses can not be accurately recovered. However, if we finetune the entire model or the person encoder, it tends to overfit to the clothing worn by the target subject. Our finetuning approach successfully retains both the person's identity and the intricate details of the input garments.

2.4. Single Stage Model vs. Cascaded

Table 4 (1st and 3rd rows) presents the FID and KID metrics on our $8,300$ triplets test set, comparing our single-stage model with the cascaded variant. Additionally, Figure 15 offers more qualitative results. While our method does not surpass the cascaded variant in terms of FID and KID scores with significant margin, the qualitative results indicate that it excels at preserving complex garment details, such as texts and logos. This observation aligns with insights from [7, 12], which suggest that FID and KID are more effective at capturing overall visual composition rather than the nuances of fine-grained visual aesthetics.

2.5. Progressive Training vs. Training from Scratch

Table 4 (2nd and 3rd rows) reveals that our progressive training strategy yields better results than training from scratch when considering FID and KID scores on our $8,300$ triplets test set. In Figure 16, we demonstrate additional qualitative results, suggesting that our progressive training approach is more effective at managing complicated garment warping.

	TryOnDiffusion [16]	Ours
SSIM ↑	0.883	**0.908**
LPIPS ↓	0.165	**0.096**

Table 5. SSIM and LPIPS scores on our $1,000$ paired test data.

2.6. Comparison on Paired Test Set

We have collected $1,000$ paired test set (not seen during training. Each pair has same person wearing the garment but under two poses). Table 5 shows that our method achieves better SSIM and LPIPS for the paired data compared to TryOnDiffusion [16]. Figure 17 shows qualitative results, where our method can better preserve intricate garment details.

2.7. Additional Qualitative Results

Figure 18 and 19 present try-on results for the dress category (denoted as I_g^{full} in the main paper). Note that our method is able to synthesize realistic folds and wrinkles in dress, well aligned with the person's pose, while preserving the intricate details of the garment. Figure 20 visualizes full images of Figure 6 in the main paper. Figure 21 provides more failure cases of our method. Finally, we provide interactive web demos for the mix and match try-on task in the supplementary material.

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 1, 2

Figure 1. **Qualitative comparison against TryOnDiffusion [16] on our 8, 300 triplets test set part one.** Our method can generate better garment details and layouts. Red boxes highlight errors of TryOnDiffusion. Please zoom in to see details.

[2] Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, et al. Pali-3 vision language models: Smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*, 2023. 1

[3] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 2

[4] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2

[5] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1

[7] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *arXiv preprint arXiv:2305.01569*, 2023. 2

[8] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. *arXiv preprint arXiv:2305.08891*, 2023. 22

[9] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 2

[10] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2231–2235, 2022. 1, 8, 9

[11] 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, 8, 9

[12] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2

[13] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 1

[14] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18359–18369, 2023. 2

[15] 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, 9

[16] Luyang Zhu, Dawei Yang, Tyler Zhu, Fitsum Reda, William Chan, Chitwan Saharia, Mohammad Norouzi, and Ira Kemelmacher-Shlizerman. Tryondiffusion: A tale of two unets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4606–4615, 2023. 1, 2, 3, 5, 6, 7, 8, 9

Figure 2. **Qualitative comparison against TryOnDiffusion [16] on our $8,300$ triplets test set part two.** Our method can generate better garment details and layouts. Red boxes highlight errors of TryOnDiffusion. Please zoom in to see details.

Figure 3. **Qualitative comparison against TryOnDiffusion [16] on our 8,300 triplets test set part three.** Our method can generate better garment details and layouts. Red boxes highlight errors of TryOnDiffusion. Please zoom in to see details.

Figure 4. **Qualitative comparison against TryOnDiffusion [16] on our 8,300 triplets test set part four.** Our method can generate better garment details and layouts. Red boxes highlight errors of TryOnDiffusion. Please zoom in to see details.

Figure 5. **Qualitative comparison against GP-VTON [15], LaDI-VTON [11] and TryOnDiffusion [16] on DressCode[10] triplets test set part one.** Ours-DressCode represents our method trained only on DressCode dataset. Red boxes highlight errors of baselines. Please zoom in to see details.

Figure 6. **Qualitative comparison against GP-VTON [15], LaDI-VTON [11] and TryOnDiffusion [16] on DressCode[10] triplets test set part two.** Ours-DressCode represents our method trained only on DressCode dataset. Red boxes highlight errors of baselines. Please zoom in to see details.

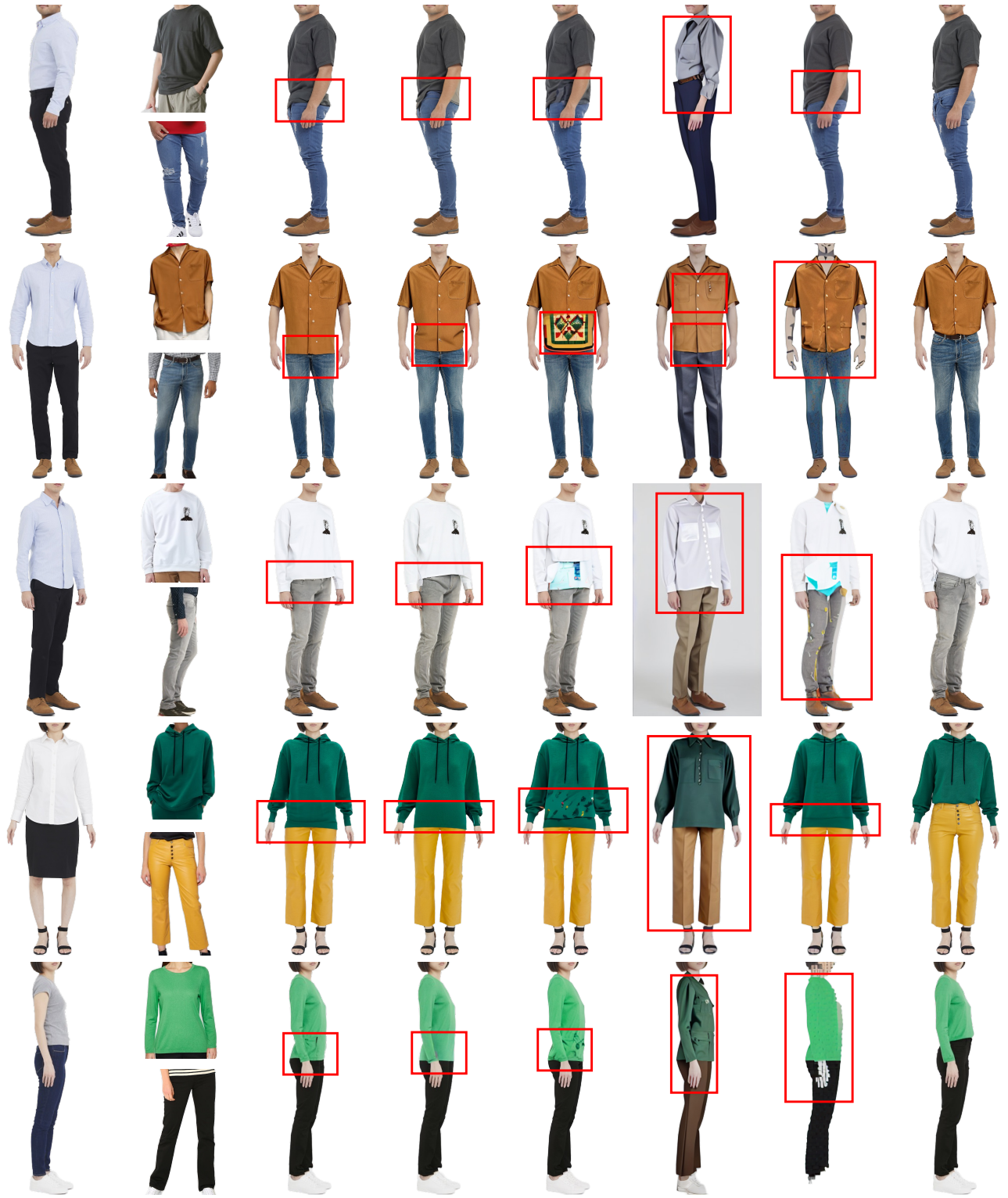

Figure 7. **Qualitative comparison for editing instruction: "tuck in the shirt".** Please zoom in to see how our method can perform the desired editing while preserving garment details. Red boxes highlight errors of baselines.

Figure 8. **Qualitative comparison for editing instruction: "tuck out the shirt".** Please zoom in to see how our method can perform the desired editing while preserving garment details. Red boxes highlight errors of baselines.

Figure 9. **Qualitative comparison for editing instruction: "roll down the sleeve".** Please zoom in to see how our method can perform the desired editing while preserving garment details. Red boxes highlight errors of baselines.

Figure 10. **Qualitative comparison for editing instruction: "roll up the sleeve".** Please zoom in to see how our method can perform the desired editing while preserving garment details. Red boxes highlight errors of baselines.

Figure 11. **Qualitative comparison for person finetuning of subject 1.** Please zoom in to see how our method can preserve both person identity and garment details. Red boxes highlight errors of baselines.

Figure 12. **Qualitative comparison for person finetuning of subject 2.** Please zoom in to see how our method can preserve both person identity and garment details. Red boxes highlight errors of baselines.

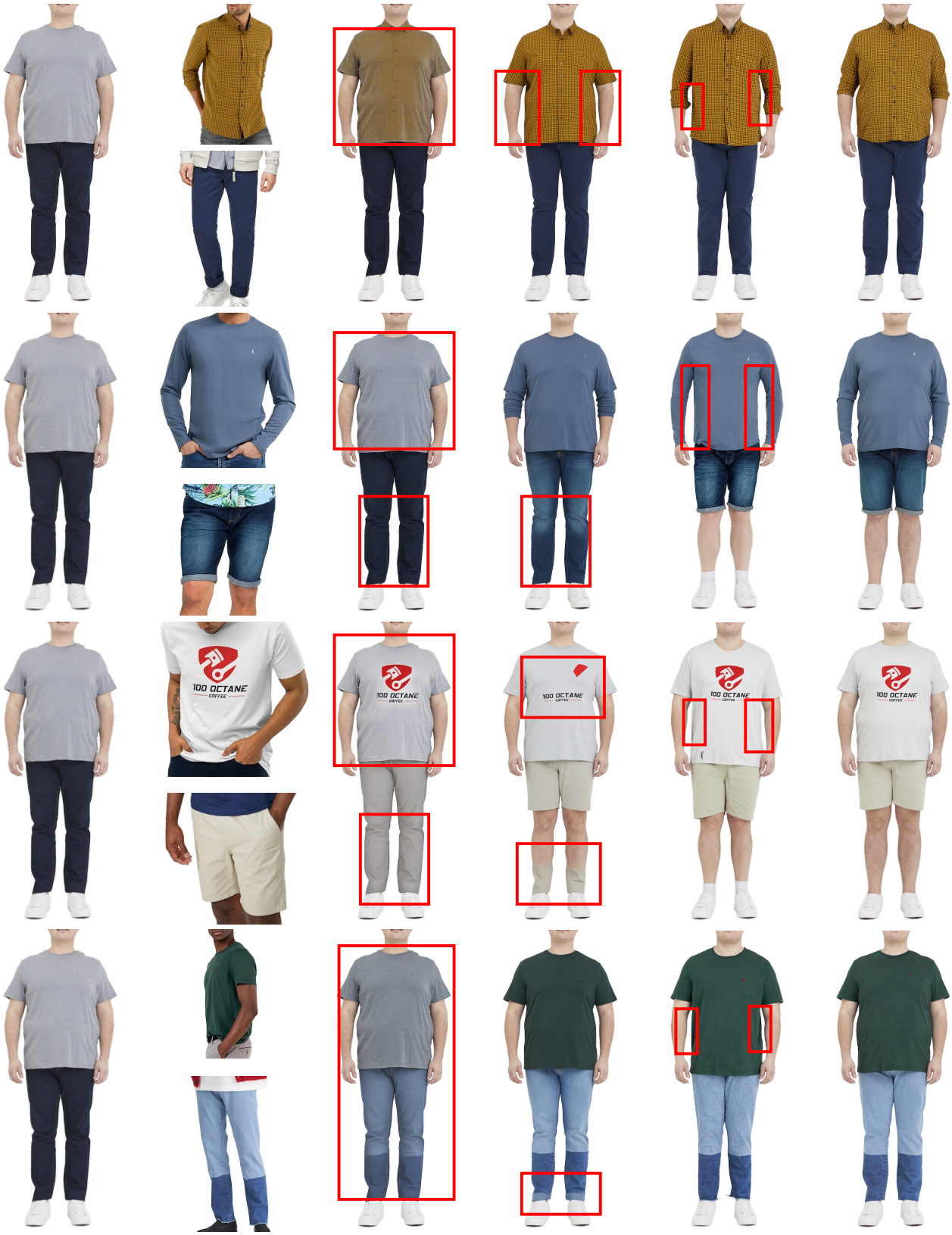

Figure 13. **Qualitative comparison for person finetuning of subject 3.** Please zoom in to see how our method can preserve both person identity and garment details. Red boxes highlight errors of baselines.

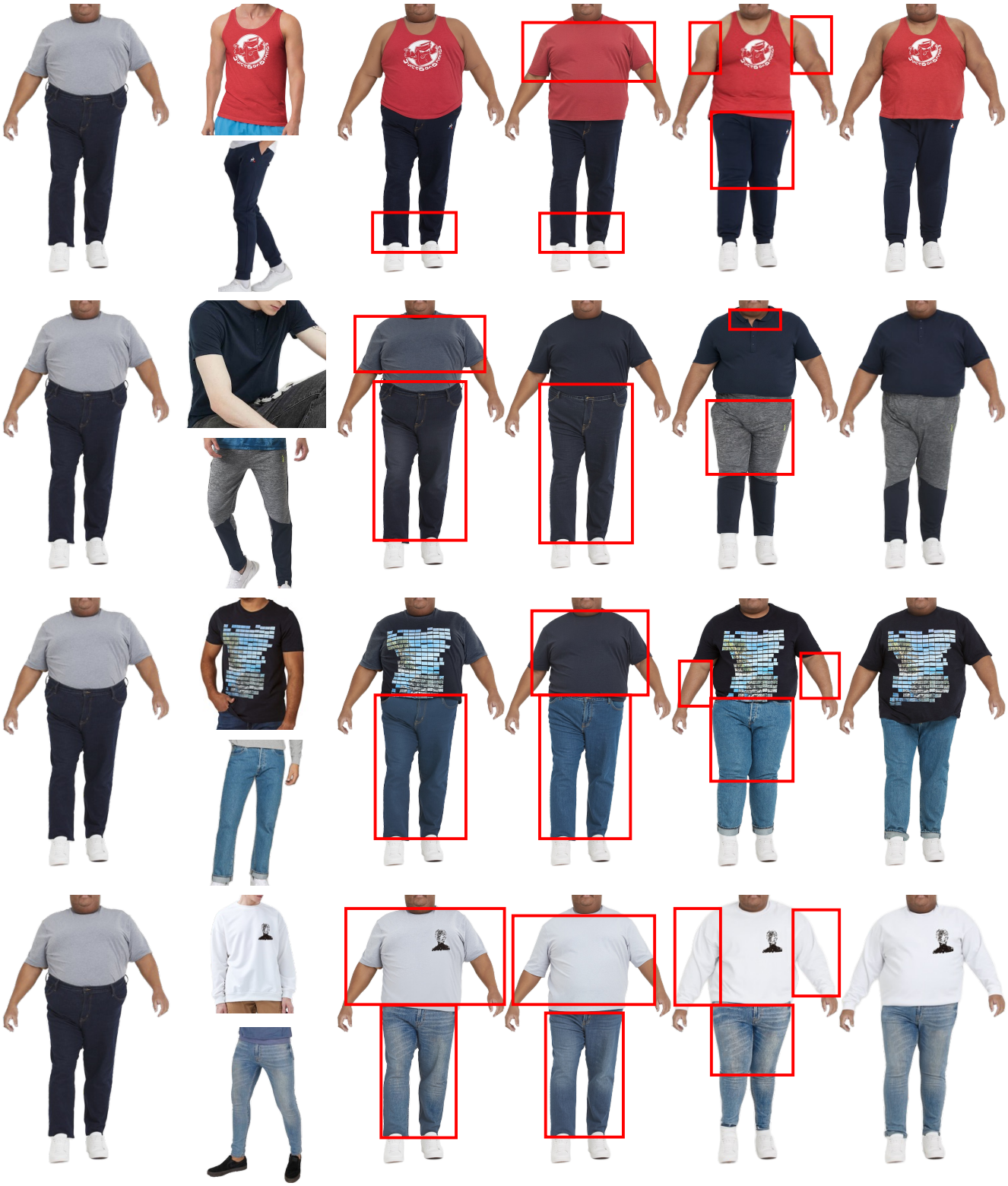

Figure 14. **Qualitative comparison for person finetuning of subject 4.** Please zoom in to see how our method can preserve both person identity and garment details. Red boxes highlight errors of baselines.

Input Person | Input Garments | Cascaded Models | Ours Single Stage | Input Zoom-in | Cascaded Model Zoom-in | Ours Single Stage Zoom-in

Figure 15. **Qualitative comparison for single stage model vs cascaded.** Our proposed single stage model can preserve fine garment details like text and logos under large pose differences. The last three columns visualize zoom-ins of red boxes for input, cascaded variant and single stage model respectively. Please zoom in to see details.

Figure 16. **Qualitative comparison for progressive training vs training from scratch.** Training from scratch can not handle complicated garment warping. Red boxes highlight errors of the training from scratch variant. Please zoom in to see details.

Figure 17. **Qualitative comparison on our** $1,000$ **paired test data.** Red boxes highlight errors of baselines. Zoom in to see details.

Figure 18. **Qualitative results for Dress VTO part one.** Our approach effectively manages complex garment warping and generates realistic wrinkles that align with the person's pose. Please zoom in to see details.

Figure 19. **Qualitative results for Dress VTO part two.** Our approach effectively manages complex garment warping and generates realistic wrinkles that align with the person's pose. Please zoom in to see details.

Figure 20. **Full images of Figure 6 in the main paper.** Please zoom in to see details.

Figure 21. **More failure cases.** Top left: our method sometimes suffers from color drift issues for very dark images, which is recognized by diffusion literature [8]. Top right: our method fails to generate valid layout for uncommon garment combinations (*e.g.* long coat and skirt). Bottom left: the model attempts to create a pocket to accommodate the occluded left hand. Bottom right: our model could generate a random inner top given "outer top open" garment layout. Additionally, it has difficulties in effectively warping small, densely packed, and irregularly distributed texture patterns.