RefTon: Reference person shot assist virtual Try-on

Supplementary Material

A. Generation of Reference Image

This section provides a detailed description of the process for generating the reference image. Many virtual try-on datasets offer the garment image and the image of the target person wearing the target garment, but they do not include the reference image, which shows the visual effect of the target garment $\boldsymbol{c}_i$ being worn by another person $\boldsymbol{p}_{\boldsymbol{c}_j}$. The generation of the reference image can be viewed as an image editing task on $\boldsymbol{p}_{\boldsymbol{c}_j}$. As discussed in Section 3.3, the reference image $\boldsymbol{r}_i$ must satisfy three key requirements.

Firstly, the reference image should faithfully preserve the details of the target garment $\boldsymbol{c}_i$. This requires the editing model to have strong consistency abilities. The *Flux-Kontext* model has a robust ability to edit the target region corresponding to the text prompt while keeping unrelated regions—such as the area not related to the garment—unchanged. Moreover, the *Flux-Kontext* model can perform precise local editing according to the text prompt, in this case, focusing on the person under the garment. Therefore, we choose the popular *Flux-Kontext* model [36] to edit the input image conditioned on the text prompt. Specifically, we add a sentence such as "keep the {target cloth} cloth unchanged" in the text prompt.

Secondly, the person in the reference image should look significantly different from the person in the target image. During training, we observed that if the reference image is too similar to the target image, the model tends to rely on a "shortcut"—directly copying from the reference image and ignoring the agnostic/person image $\boldsymbol{a}_i$/$\boldsymbol{p}_i$ and the cloth $\boldsymbol{c}_i$. To avoid this, we ensure that the person in the reference image differs from the target image to better showcase the visual effect of the clothing when worn, rather than focusing on the appearance of the person themselves. To achieve this, we utilize the Text-to-Image (T2I) capabilities of the *Flux-Kontext* model. We extract an accurate description of the person's appearance in the target image (e.g., "The model has an East Asian appearance, with light skin, long black hair, and a neutral expression...") and pass it as the **negative prompt** to the T2I model. In contrast, we provide an opposite description (e.g., "The model has an African appearance, with dark skin, short yellow hair, and a cheerful expression...") as the **positive prompt** to guide the editing of the target image, as shown in Figure 4(b).

However, extracting descriptions for each image manually is labor-intensive. To address this, we use a vision-language model (VLM) such as *Qwen2.5-VL* to automatically generate the description and its opposite. Specifically, we pass the target image $\boldsymbol{p}_{i,\boldsymbol{c}_i}$ to the VLM and provide a prompt like "Start with Positive: describe only the model's race, skin, hair, eyes, and expression, then give the opposite in one sentence with Negative: changing those traits without 'not' or clothing." The generated description and its opposite are then fed into the negative and positive prompt encoders of the T2I model to edit the target image, as shown in Figure 4(a). In this way, we automate the generation of the reference image $\boldsymbol{r}_i$ by editing the target image.

Thirdly, after extracting the description and opposite description of the person's appearance, we introduce more diversity into the reference image by varying the non-target garments and actions of the human in the target image. This can be achieved through the image editing model by adding descriptions related to non-target garments and actions. We provide a description bank containing candidate descriptions for outfits and actions across three scenarios: the person in the image is wearing a dress, an upper-body garment, or a lower-body garment. This ensures that the description of the editable garment differs from the target garment $\boldsymbol{c}_i$. Furthermore, the outfit descriptions also include accessories such as glasses, wristwatches, and bracelets to increase diversity. These descriptions, along with the actions and outfit details, are concatenated into positive prompts and passed to the T5 text encoder, as shown in Figure 4.

Fig S1 illustrates selected text prompts from the prompt description bank used for reference image generation. Furthermore, Fig S2 exhibits a selection of the resulting reference data samples.

B. Further Quantitative Evaluation on High Resolution

Here we provide more results of the generated image. Table S1 summarizes the detailed quantitative evaluation on the DressCode dataset. Our method (RefTON) outperforms all baselines, delivering higher try-on quality and achieving strong consistency with the target person's pose and body structure. Integrating reference images ("+R") further enhances the results, establishing a new state-of-the-art. Importantly, even in the mask-free setting—without agnostic masks or additional inputs—our method correctly preserves garment styles (e.g., clothing length and design) and maintains high pose alignment, while achieving accuracy comparable to or surpassing prior baselines, demonstrating robustness and practicality.

To further evaluate our method under high-resolution settings, Table S2 reports quantitative results on both VITON-HD and DressCode at a resolution of 1024. Across the paired and unpaired protocols, RefTON and RefTON+R

"Dresses":

[{"action": "She steps forward lightly, one arm swinging gently while the other rests on her waist, ",
"outfit": "wearing a delicate silver bracelet and black high heels. Keep the dress exactly as it is." },

{ "action": "She turns smoothly on one foot, arms extended outward gracefully, ",
"outfit": "wearing a thin gold ring and beige high heels. Keep the dress exactly as it is." },
… …]

"Upper Body":

{"action": "The model shifts shoulders with a playful tilt, arms lifting lightly while one foot slides outward.",
"outfit": "The model is wearing navy cotton trousers with a plain texture, dark leather sandals, and a leather bracelet. Keep the upper body clothing exactly as it is, only change the lower body clothes or shoes." },

{"action": "The model rotates the torso in motion, one hand extending forward at chest height, the other resting at the side, legs following the twist.",
"outfit": "The model is wearing black denim trousers with a matte finish, white canvas sneakers, and a wristwatch. Keep the upper body clothing exactly as it is, only change the lower body clothes or shoes."},
… …]

"Lower Body":

[{"action": "She steps forward lightly, one arm swinging gently while the other rests on her waist, ",
"outfit": "wearing a delicate silver bracelet and black high heels. Keep the dress exactly as it is." },

{ "action": "She turns smoothly on one foot, arms extended outward gracefully, ",
"outfit": "wearing a thin gold ring and beige high heels. Keep the dress exactly as it is." },
… …]

Figure S1. Sample text prompts from the Outfit and Action Description Bank. To ensure the model edits only the person while preserving the target clothing, we assign different outfits and action description categories to different clothing inputs.

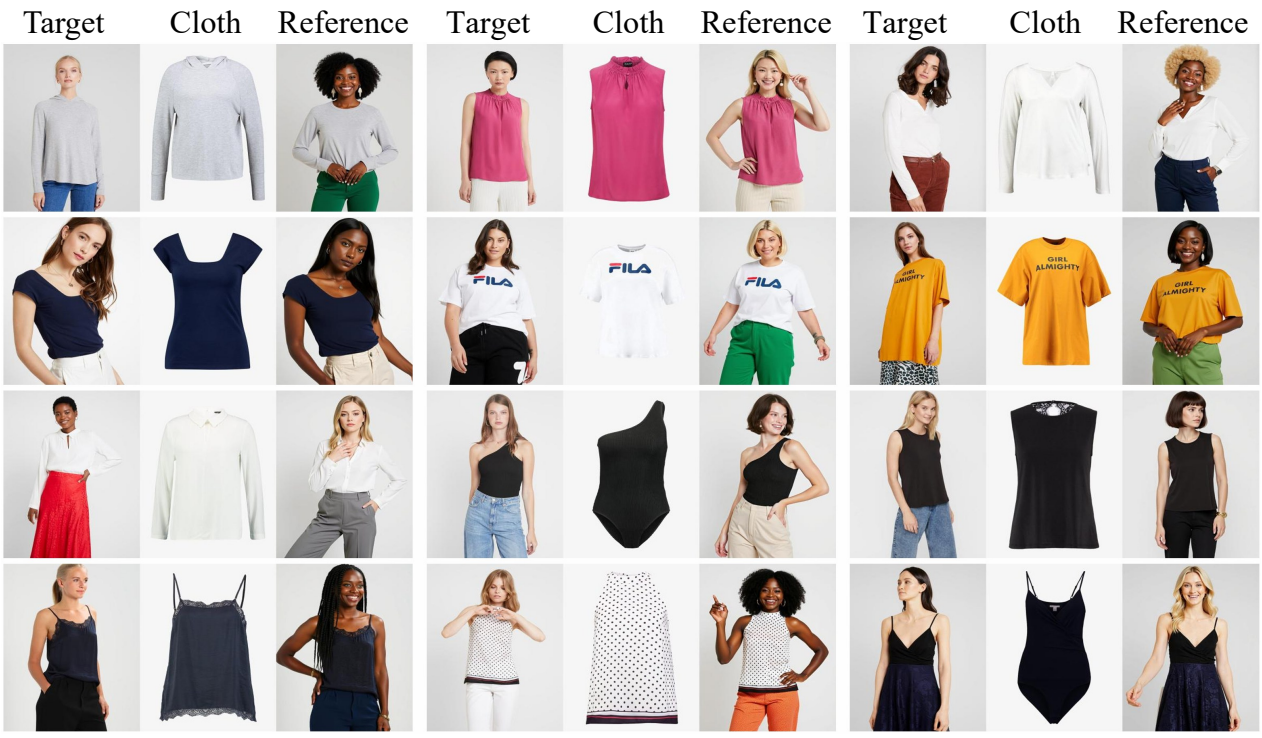

Figure S2. Sample reference images generated by our reference data generation pipeline. The editing model takes the target person's image as input and synthesizes corresponding reference images, while preserving the garment's appearance to match the cloth.

Table S1. Quantitative results on three subsets of the DressCode dataset [44]: upper body, lower body, and dresses. The best and second-best results are highlighted in **bold** and underline, respectively. The symbol "*" denotes results reported in prior work, while "+R" indicates results with reference image input, and "MF" refers to mask-free input images. The subscripts p and s denote specific evaluation metrics for precision and recall, respectively.

Method	Upper-body				Lower-body				Dresses			
	FID$_p$↓	**KID**$_p$↓	**FID**$_u$↓	**KID**$_u$↓	**FID**$_p$↓	**KID**$_p$↓	**FID**$_u$↓	**KID**$_u$↓	**FID**$_p$↓	**KID**$_p$↓	**FID**$_u$↓	**KID**$_u$↓
CAT-DM* [63]	9.85	2.38	12.62	1.89	10.25	1.81	14.83	2.82	10.71	2.02	14.30	3.36
OOTDiffusion* [60]	11.03	0.29	–	–	9.72	0.64	–	–	10.65	0.54	–	–
PromptDresser* [31]	11.00	0.74	–	–	12.55	1.46	–	–	11.09	1.10	–	–
RefTON(Ours)	7.62	1.10	11.13	0.98	7.60	1.38	13.07	2.11	7.32	1.30	11.56	1.98
RefTON+R(Ours)	**6.39**	**0.85**	**11.08**	**0.87**	**6.61**	**1.05**	12.56	**1.67**	**6.09**	**1.16**	11.16	1.72
RefTON/MF(Ours)	8.37	1.43	11.20	1.11	8.79	1.51	**12.50**	1.83	7.36	1.33	10.73	1.41
RefTON+R/MF(Ours)	7.20	1.18	11.53	1.12	7.85	1.21	12.74	2.05	6.24	1.20	**10.05**	**1.30**

Table S2. Quantitative comparison across VITON-HD [8] and DressCode [44] at a resolution of 1024. The best and second best results are shown in **bold** and underline. "+R" denotes the use of reference images, and "MF" indicates mask-free inputs. Subscripts p and u represent the *paired* and *unpaired* test settings, respectively. Unless otherwise specified, the same notations carry the same meanings throughout the figures and tables in this paper.

Method	VITON-HD						DressCode					
	LPIPS$_p$↓	**SSIM**$_p$↑	**FID**$_p$↓	**KID**$_p$↓	**FID**$_u$↓	**KID**$_u$↓	**LPIPS**$_p$↓	**SSIM**$_p$↑	**FID**$_p$↓	**KID**$_p$↓	**FID**$_u$↓	**KID**$_u$↓
Mask-based setting												
RefTON (Ours)	0.079	0.870	5.96	1.05	**8.91**	**1.15**	0.056	0.899	3.28	0.76	4.84	0.83
RefTON+R (Ours)	**0.072**	**0.873**	**5.25**	**0.97**	9.10	1.41	**0.052**	**0.902**	**2.84**	**0.65**	**4.73**	**0.76**
Mask-Free setting												
RefTON/MF (Ours)	0.068	**0.880**	5.02	0.85	**8.87**	**1.05**	**0.028**	**0.956**	**1.03**	**0.19**	**4.24**	**0.59**
RefTON+R/MF (Ours)	**0.067**	0.875	**4.73**	**0.71**	8.98	1.22	0.030	0.953	1.15	0.25	4.41	0.69

consistently achieve high performance on nearly all metrics. Notably, the mask-free variants (RefTON/MF and RefTON+R/MF) deliver particularly strong results. These results demonstrate that our framework scales effectively to high-resolution synthesis and remains robust across diverse virtual try-on settings.

C. Additional Qualitative Results

In this section, we provide extensive qualitative visualizations to further demonstrate the robustness, generalization ability, and high-fidelity performance of our method across different datasets, clothing categories, and evaluation settings.

We present qualitative comparisons with methods that require additional text inputs, and explicitly annotate each method's required inputs in Fig S3(a). Our method achieves the best visual quality with the fewest required inputs. Moreover, adding a reference image further improves intricate details (e.g., lace, transparency, and texture), highlighting the advantage of visual reference. Figure S3(b) shows qualitative results on the StreetTryOn dataset under various settings. Compared with the baselines, our model produces more faithful try-on details while better preserving the person's pose and the background. We also provide qualitative comparisons under the mask-free setting, as shown in Fig. S3(c). Our model demonstrates the strongest mask-free virtual try-on capability among all compared methods.

As shown in Fig. S4, S5, S6, S7, and S8, our approach consistently preserves garment details, structural correctness, and texture realism under both paired and unpaired scenarios, with or without mask-free (MF) inputs. Moreover, as illustrated in Fig.S9, even in challenging in-the-wild conditions, our model exhibits strong robustness—maintaining accurate body pose, preserving background integrity, and producing stable try-on results without introducing artifacts or unintended changes.

Complex Patterns (VITON-HD). As shown in Fig. S4, our model faithfully reproduces complex and fine-grained clothing patterns. Even for garments with dense textures, irregular motifs, or high-frequency visual elements, the generated results retain clear, sharp, and recognizable patterns with minimal distortion. The strong pattern-preservation ability highlights the effectiveness of our approach in capturing both global appearance and subtle local details.

Complex Structures (VITON-HD). Fig. S5 further illustrates that our method handles garments with challenging

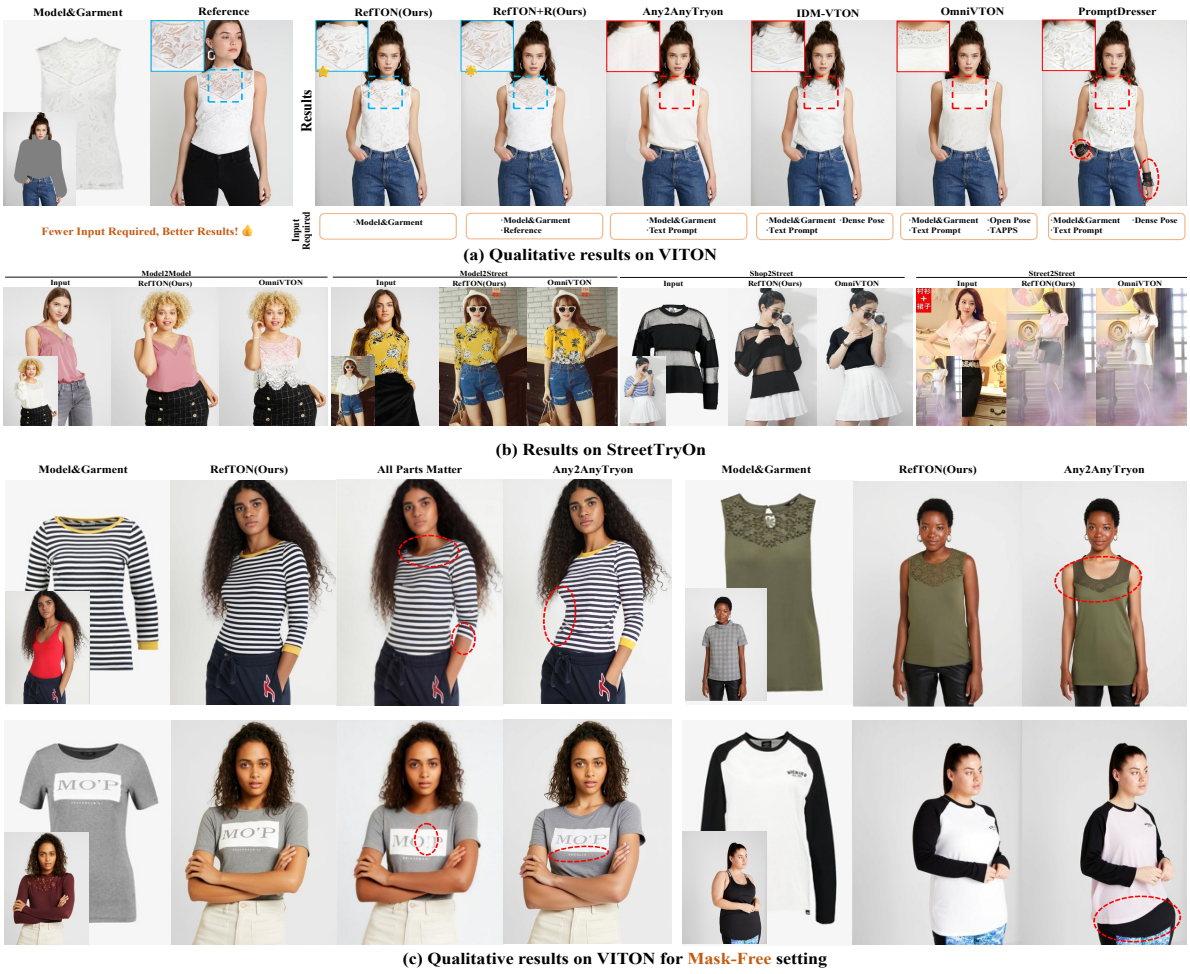

Figure S3. (a)We gain better results with fewer inputs required on VITON-HD. (b)StreetTryOn qualitative results (in-the-wild & P2P): our method yields better results.(c) Our model shows stronger mask-free capability.

structural designs, such as multi-layered regions, unique silhouettes, or uncommon shapes. The generated try-on results maintain correct garment geometry, coherent contours, and physically plausible spatial arrangements. This demonstrates that our framework models structural priors robustly, enabling accurate synthesis even under significant variations in shape.

DressCode Upper-body, Lower-body, and Dress Subsets. As shown in Figs. S6, S7, and S8, our approach performs consistently well across the three DressCode subsets. In the unpaired and mask-free settings, our model successfully preserves fabric materials, shading, and texture characteristics while achieving realistic alignment between the garment and human body. Across diverse clothing types—including tops, pants, skirts, and full-body dresses—the synthesized results maintain stable structure, smooth boundaries, and visually coherent integration, demonstrating strong generalization and robustness.

In-the-Wild results. In addition to controlled benchmark evaluations, RefTON demonstrates strong robustness and generalization in challenging *in-the-wild* scenarios. As shown in Fig. S9, it produces high-quality try-on results under diverse poses, lighting conditions, and cluttered backgrounds. Our **mask-free** pipeline directly transfers garments without human parsing masks or pose estimators, while preserving body pose, global structure, and identity cues. Moreover, **additional-reference try-on** further improves garment geometry, texture details, and overall realism. RefTON also maintains strong background consistency and avoids unnecessary changes to non-garment regions, making it reliable for real-world applications.

Figure S4. **Qualitative paired results in VITON-HD dataset with complex patterns on clothes.** "reference" denotes that a reference image is provided.

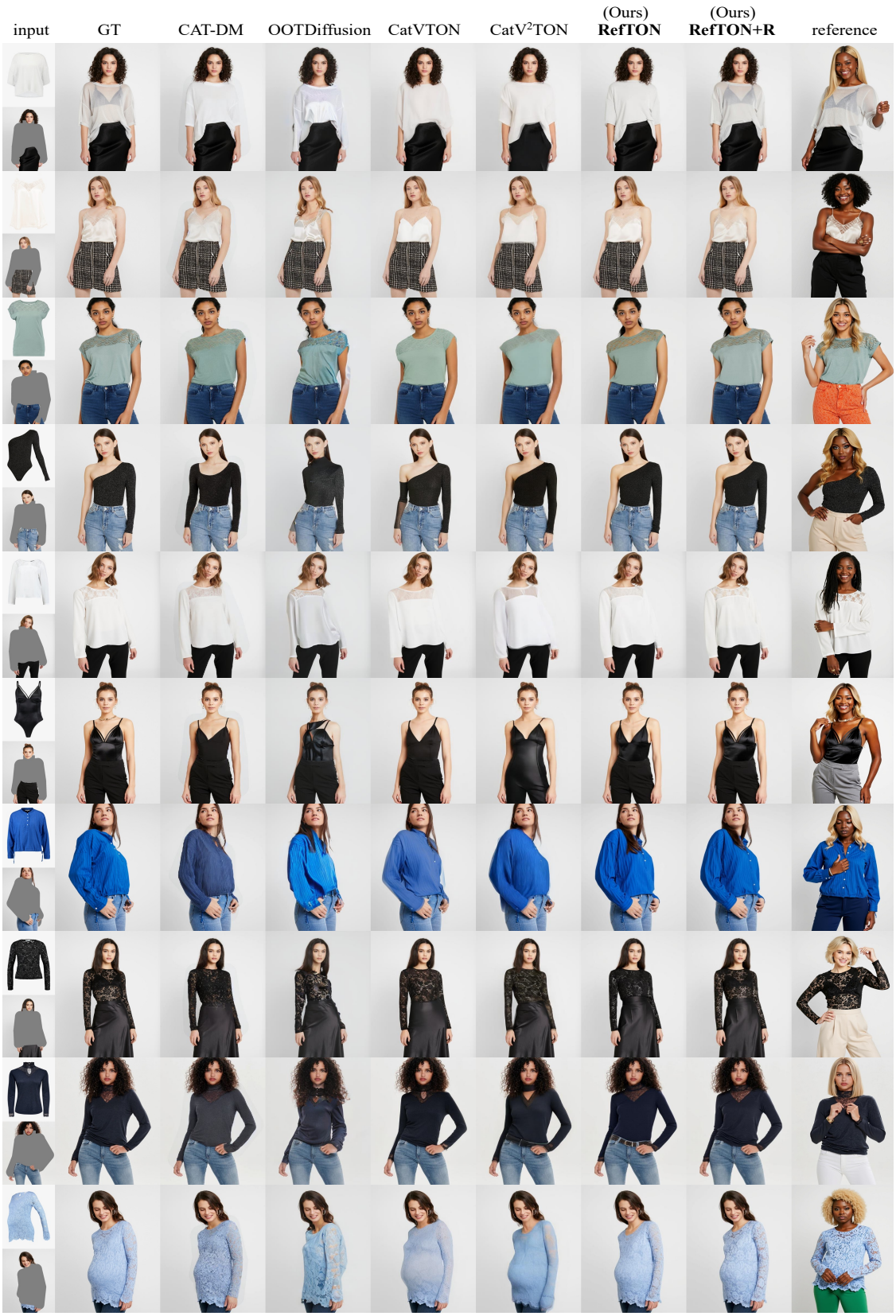

Figure S5. **Qualitative paired results in VITON-HD dataset with complex structure on clothes.** "reference" denotes that a reference image is provided.

Figure S6. **Qualitative results of upper-body sub-set in Dresscode dataset unpaired setting.** "reference" denotes that a reference image is provided, while "MF" indicates mask-free inputs using the original person image instead of a masked agnostic image.

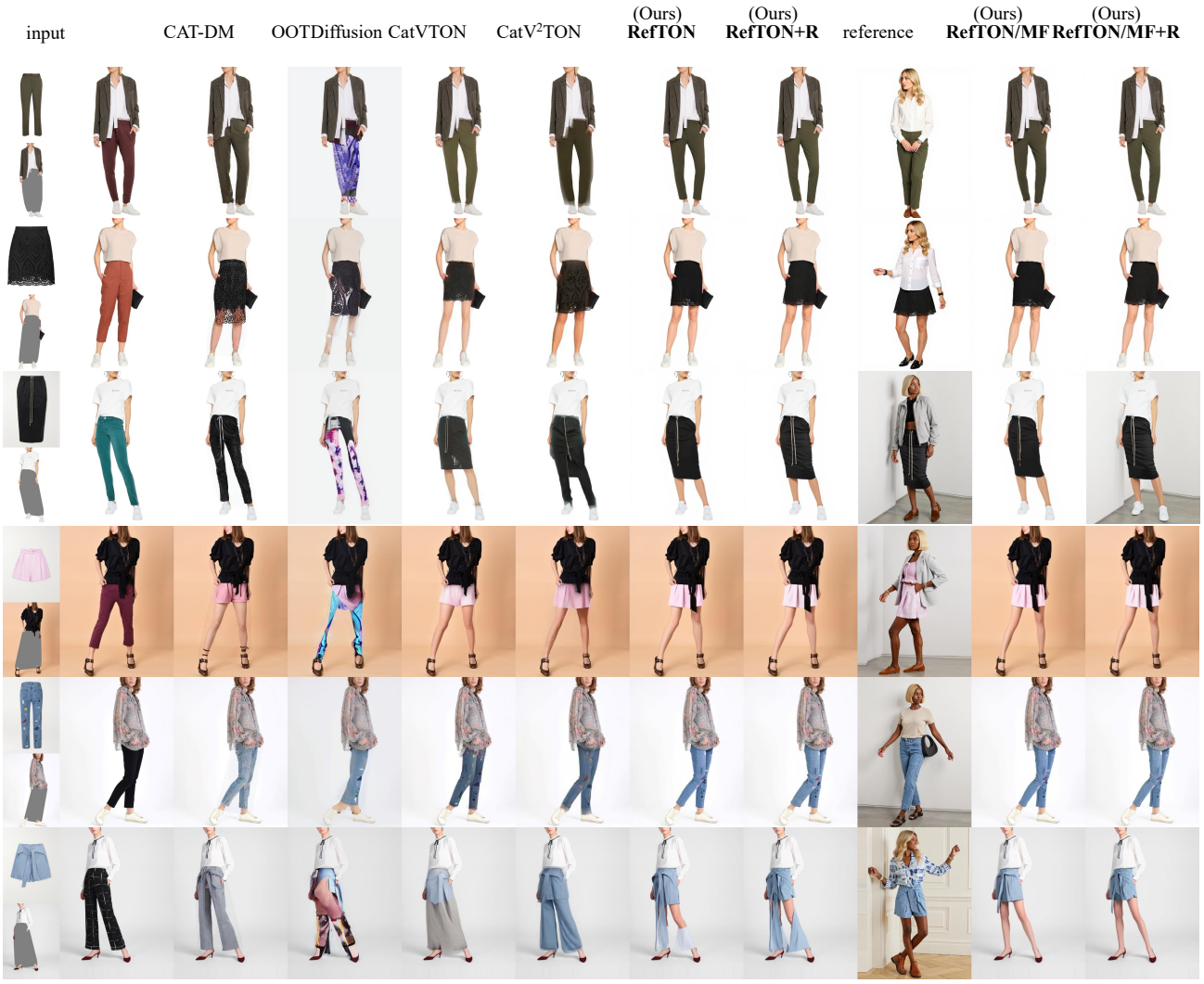

Figure S7. **Qualitative results of lower-body sub-set in Dresscode dataset unpaired setting.** "reference" denotes that a reference image is provided, while "MF" indicates mask-free inputs using the original person image instead of a masked agnostic image.

Figure S8. **Qualitative results of dresses sub-set in Dresscode dataset unpaired setting.** "reference" denotes that a reference image is provided, while "MF" indicates mask-free inputs using the original person image instead of a masked agnostic image.

Figure S9. **In-the-wild try-on results produced by our RefTON model.** The first row demonstrates our **mask-free try-on** capability, where the garment is transferred directly to the target person without requiring human parsing masks or pose estimation. The second row shows our **additional-reference try-on** mode, in which extra visual references are incorporated to enhance structural accuracy, texture fidelity, and overall realism.