OmniVTON: Training-Free Universal Virtual Try-On – Supplementary Materials –

Zhaotong Yang[1], Yuhui Li[1], Shengfeng He[2], Xinzhe Li[1], Yangyang Xu[3], Junyu Dong[1], Yong Du[1]*
[1]Ocean University of China,
[2]Singapore Management University, [3]Harbin Institute of Technology (Shenzhen)

1. Details and Discussion

1.1. Implementation Details

All experiments were conducted using PyTorch 2.1.1 on a NVIDIA GeForce RTX 3090 GPU. We adopted Stable Diffusion v2 [6] as the base model, retaining default hyperparameter configurations. For both Pseudo-Person Image Generation and Garment-Infused Image Inpainting, we employed the standard DDIM sampler for deterministic inference with 50 time steps. For Spetral Pose Injection, we set the standard deviation τ of the Gaussian mask to 0.1.

During the garment morphing stage, we implemented distinct region segmentation strategies for different garment categories: 1) Upper garments underwent five-region processing (left and right upper arms, left and right lower arms, and torso regions); 2) Lower garments were similarly decoupled into five regions (left and right upper legs, left and right lower legs, and hip-above regions); 3) Dresses were segmented into upper and lower garment sections for separate processing. The agnostic and clothing masks are provided by the dataset. In practical applications, SAM [3] can be used to obtain the mask corresponding to the user input image.

1.2. Text Prompts Acquisition

Here, we describe the process of acquiring text prompts and examine their impact. Specifically, we convert images into text using the CLIP Interrogator [5], where the generated descriptions consist of a core caption and auxiliary modifier terms. The core caption directly describes the image content, while the auxiliary terms are selected based on cosine similarity between garment features and text embeddings from four predefined datasets: artists, mediums, movements, and flavors.

To verify the importance of text prompts in virtual try-on tasks, we conducted a controlled analysis using a generic prompt ("a person wearing an upper garment"). As shown in Fig. 1, more detailed text prompts lead to try-on results with enhanced identity consistency, highlighting the crucial role of precise textual descriptions in controlling the quality of generation.

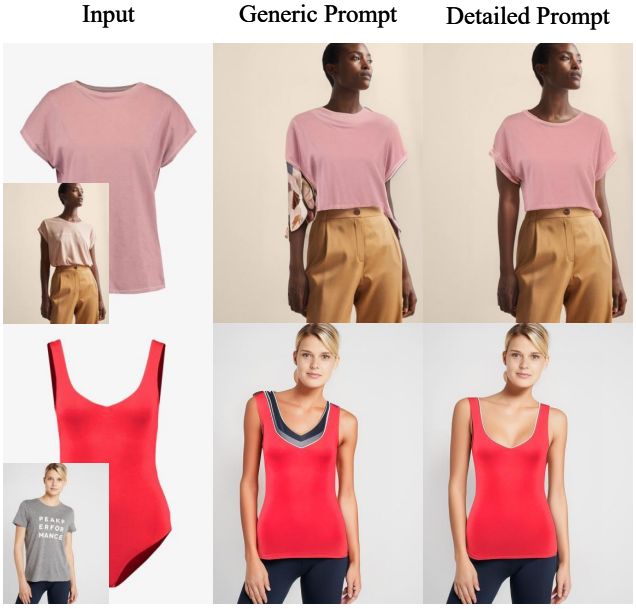

Figure 1. Influence of different text prompts.

Method	$FID_u\downarrow$	$FID_p\downarrow$	$SSIM_p\uparrow$	$LPIPS_p\downarrow$
OmniVTON	**9.621**	**7.758**	0.832	**0.145**
w/o semantic parsing	13.705	11.930	0.817	0.170
w/o I_c-path attention modulation	9.808	7.939	0.831	0.149
w/o high-frequency noise	15.817	14.558	0.836	0.182
w/o SPI + w/ average noise	12.402	10.650	**0.849**	0.151
w/o SPI + w/ ControlNet	10.873	9.016	0.818	0.168

Table 1. More ablation studies of different components.

1.3. Additional Ablation Analysis

To demonstrate the rationale behind our component design, we conducted additional ablation experiments. First, for SGM, the role of semantic parsing is to perform pixel-

*Corresponding author (csyongdu@ouc.edu.cn).

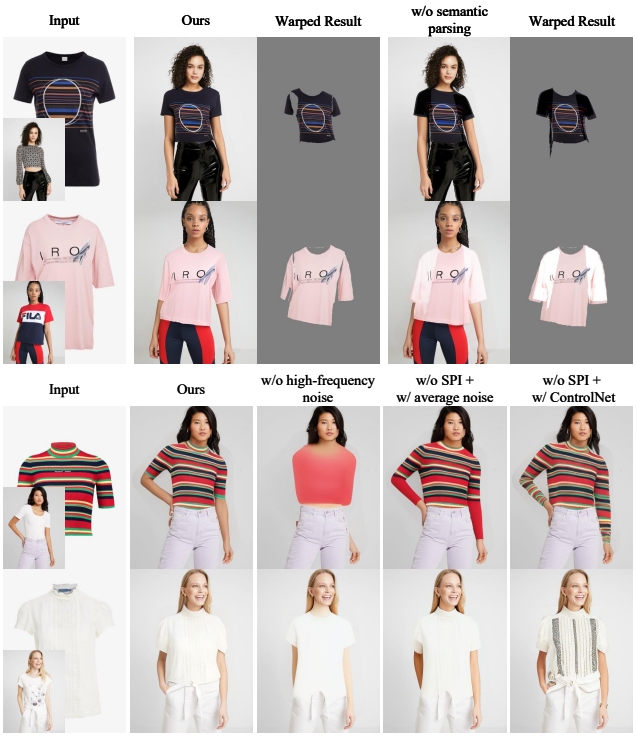

Figure 2. Qualitative results of additional ablation analysis.

level segmentation on skeleton-divided semantic regions, enabling multi-part decoupling. As shown in the upper part of Fig. 2, relying solely on bounding box-based segmentations, without semantic parsing, for localized transformations leads to erroneous morphing and part overlap, significantly degrading the quality of the try-on results. The quantitative comparison of the "w/o semantic parsing" setting in Tab. 1 strongly reinforces the necessity of this component. Secondly, the "w/o I_c-path attention modulation" setting involves replacing the attention modulation in Eq. (8) of the main paper with the original self-attention mechanism, resulting in noticeable degradation across all evaluation metrics, thus validating the effectiveness of bidirectional semantic context interaction.

For SPI, the lower part of Tab. 1 and Fig. 2 present both quantitative and qualitative results for different variants. The "w/o high-frequency noise" variant retains only the low-frequency components of inversion noise, yet the absence of high-frequency noise leads to overly smoothed results. The "w/o SPI + w/ average noise" variant averages random noise and inversion noise as the initial noise. Compared with the "w/o high-frequency noise" variant, the introduction of random noise significantly improves perceptual quality. However, due to the lack of frequency-domain decoupling, this variant enhances performance in paired settings but fails to suppress source garment texture interference from inversion noise in unpaired settings, causing performance degradation. Furthermore, comparative experiments with ControlNet [7]-based skeleton-conditioned injection demonstrate that OmniVTON effectively overcomes the inherent bias of diffusion models in handling multiple conditions by decoupling garment and pose constraints, leading to improved try-on results.

τ	$\text{FID}_u\downarrow$	$\text{FID}_p\downarrow$	$\text{SSIM}_p\uparrow$	$\text{LPIPS}_p\downarrow$
0.01	9.941	8.185	0.823	0.160
0.05	9.620	8.033	0.829	0.153
0.1	9.621	7.758	0.832	0.145
0.3	10.056	8.150	0.842	0.140
0.5	11.330	9.422	0.852	0.138

Table 2. Sensitivity analysis of cutoff frequency τ on VITON-HD.

In Tab. 2, we provide additional analysis on the sensitivity of the cutoff frequency τ. When τ is too small, it suppresses low-frequency pose information, limiting SSIM. As τ increases, metrics generally improve; however, if τ becomes too large, it preserves excessive high-frequency details, which harms realism and worsens FID. Setting $\tau = 0.1$ balances pose consistency and visual fidelity.

1.4. Inference Cost

As shown in the upper part of Tab. 3, we compare the inference costs of OmniVTON with three state-of-the-art methods. The results show that OmniVTON achieves the lowest memory consumption, outperforms Cross-Image in inference speed, and performs comparably to TIGIC and IDM-VTON, all while maintaining optimal performance. The lower part of the table further presents a module-wise breakdown of inference times. Notably, under the Non-Shop-to-X setting, removing the pseudo-person generation step leads to a sharp reduction in the runtime of the SGM module, from 6.61s to just 0.14s, thereby reducing the overall inference time to 9.82 seconds and further highlighting OmniVTON's strong potential for real-world deployment.

1.5. User Study

We validate the effectiveness of our method through a rigorously designed user study, establishing a systematic evaluation framework across three benchmark datasets: VITON-HD [1], DressCode [4], and StreetTryOn [2]. The experiment involved 100 volunteers, each participating in a visual evaluation questionnaire containing 100 comparative sample groups. Specifically, the VITON-HD dataset includes 20 test sample groups, the DressCode dataset covers 40 sample groups across three garment categories (upper, lower, dresses), and the StreetTryOn benchmark allocates the remaining 40 sample groups with a scenario-balanced distribution. Each task in the questionnaire asks, "Which method generates more realistic and accurate images?" with randomized option ordering to ensure unbiased results. As shown in Fig. 3, our method demonstrates significant supe-

Time / Memory	Training-free						Training	
	OmniVTON		TIGIC		Cross-Image		IDM-VTON	
	16.29s	11,542MB	13.87s	23,578MB	41.49s	15,748MB	11.87s	17,936MB

OmniVTON	$FID_u \downarrow$	SGM Time (s)	SPI Time (s)	CBS Time (s)
	9.621	6.61s	3.60s	6.08s

Table 3. Runtime and memory comparison on VITON-HD.

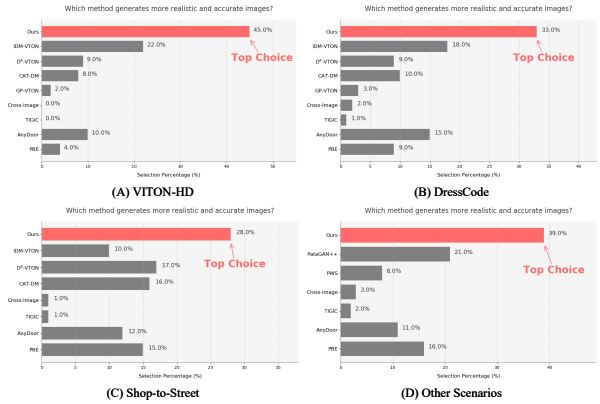

Figure 3. User study on the VITON-HD dataset [1], DressCode dataset [4] and StreetTryOn benchmark [2].

riority across all benchmarks.

1.6. Failure Case Visualizations

We present several failure cases of OmniVTON in Fig. 4. As discussed in the main paper, our method encounters challenges in handling high-density crowds and targets with minimal visible body regions. These limitations primarily stem from OmniVTON's partial reliance on pre-trained modules such as OpenPose and TAPPS, whose predictions can be unreliable under such extreme conditions. Such observations point to a promising direction for future work towards more robust and adaptable universal virtual try-on systems.

2. Additional Visual Results

2.1. Visual Comparisons with SOTAs

Fig. 5 and Fig. 6 present supplementary visual comparisons between OmniVTON and baseline methods on the VITON-HD and DressCode datasets, respectively. While Fig. 7, Fig. 8, Fig. 9, and Fig. 10 showcase detailed visualized results of different methods across four scenarios in the StreetTryOn benchmark.

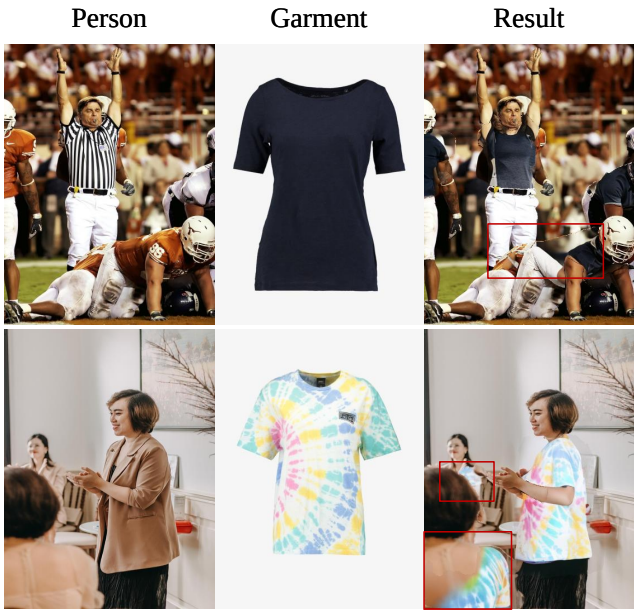

Figure 4. Failure cases of our method.

2.2. More Try-on Results

As shown in Fig. 11, we further showcase various garment-model combinations, including virtual try-on results for lower-body garments and dresses under the Shop-to-Street scenario. This highlights OmniVTON's ability to overcome the technical barriers that previously limited the performance of StreetTryOn in this task.

References

[1] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *CVPR*, pages 14131–14140, 2021. 2, 3

[2] Aiyu Cui, Jay Mahajan, Viraj Shah, Preeti Gomathinayagam, Chang Liu, and Svetlana Lazebnik. Street tryon: Learning in-the-wild virtual try-on from unpaired person images. In *WACV*, pages 8235–8239, 2025. 2, 3

[3] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer White-

Figure 5. Qualitative comparison on the VITON-HD dataset.

Figure 6. Qualitative comparison on the DressCode dataset.

Figure 7. Qualitative comparison for Shop-to-Street scenario on the StreetTryOn benchmark.

head, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 1

[4] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: High-resolution multi-category virtual try-on. In *CVPR*, pages 2231–2235, 2022. 2, 3

[5] pharmapsychotic. Clip-interrogator. https://github.com/pharmapsychotic/clip-interrogator, 2023. 1

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1

[7] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 2

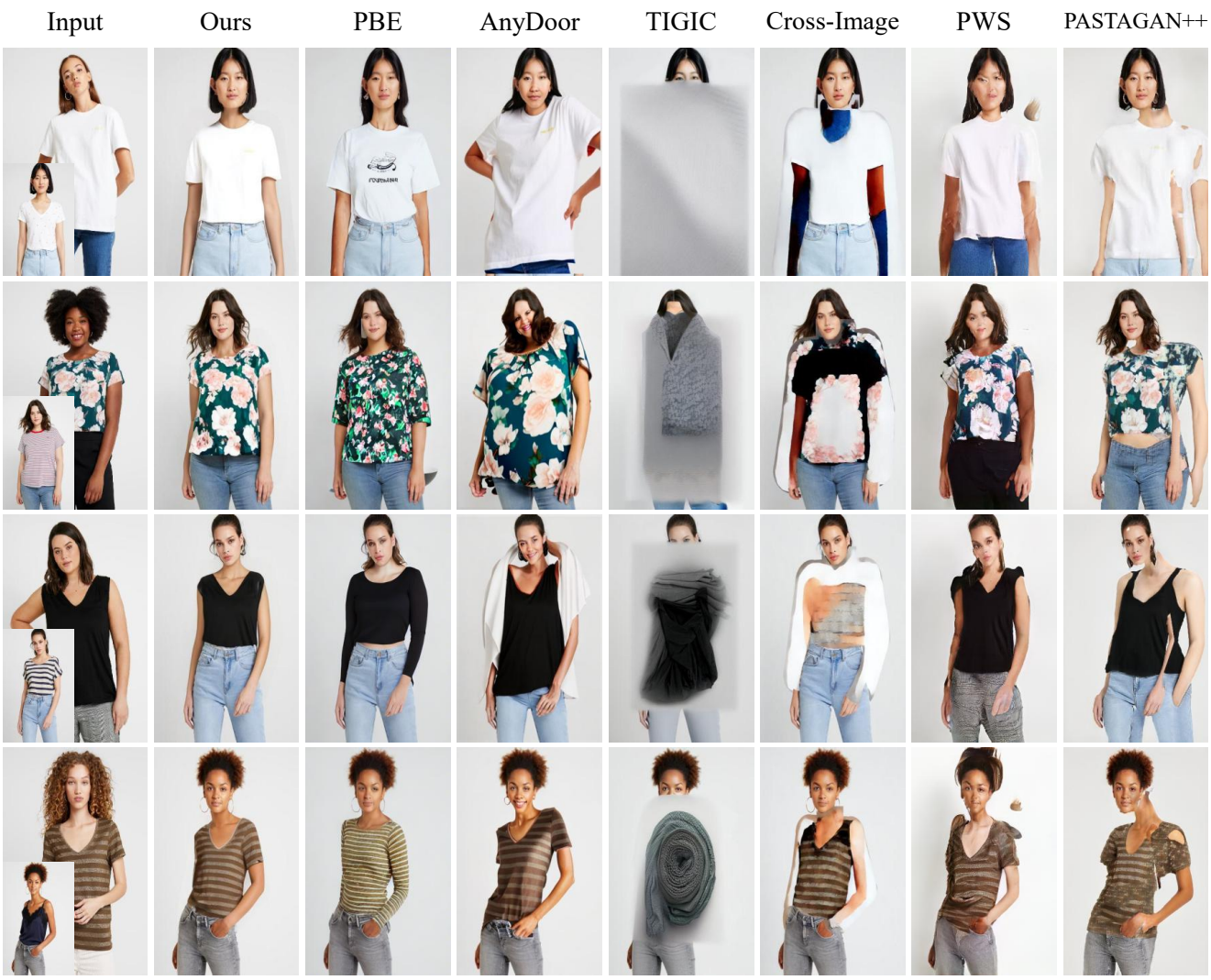

Figure 8. Qualitative comparison for Model-to-Model scenario on the StreetTryOn benchmark.

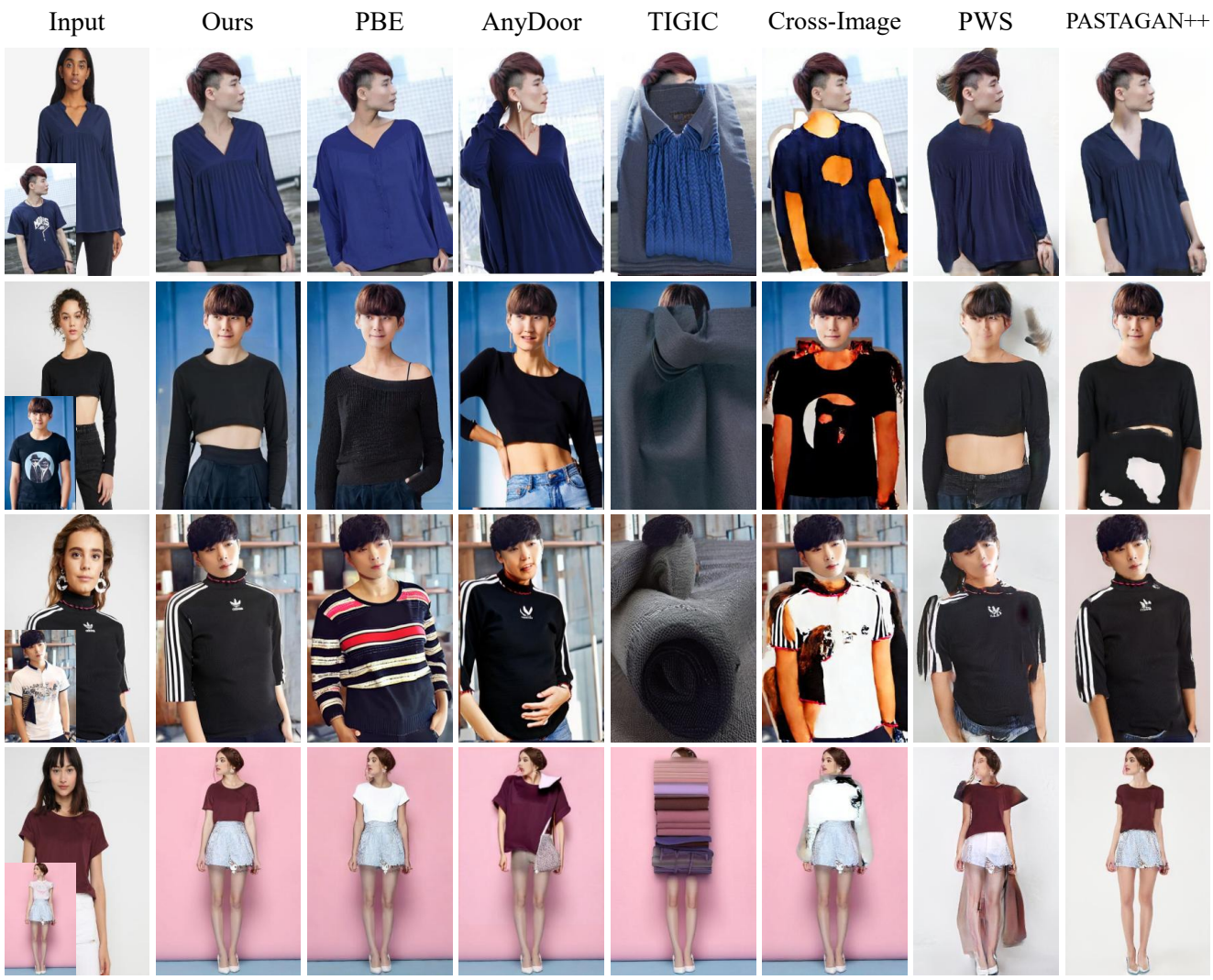

Figure 9. Qualitative comparison for Model-to-Street scenario on the StreetTryOn benchmark.

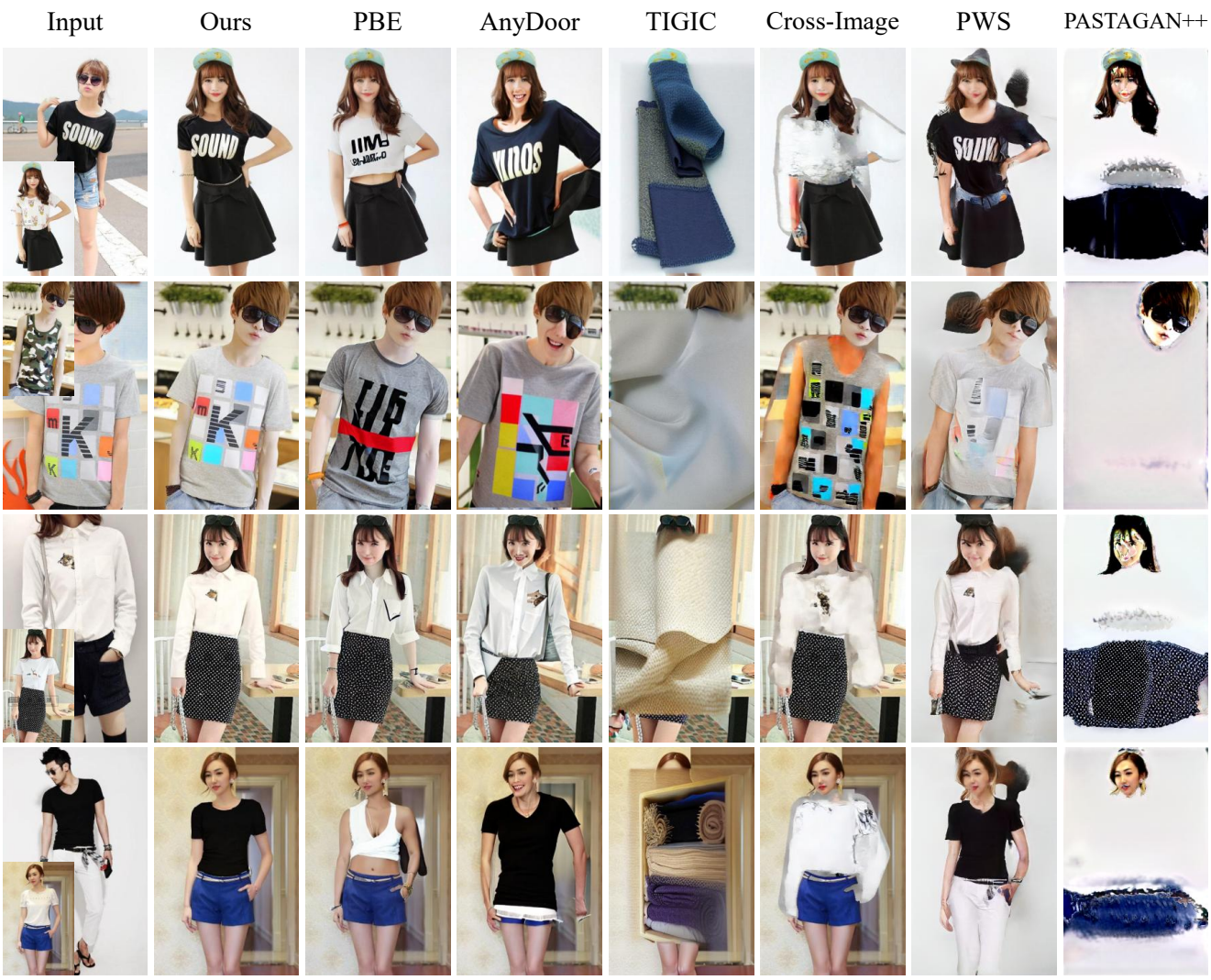

Figure 10. Qualitative comparison for Street-to-Street scenario on the StreetTryOn benchmark.

Figure 11. More try-on results of OmniVTON across various clothing types and scenarios.