Contents

1 Method Details

1.1 High Quality Pseudo Data Details

In this section, we provide additional details on creating High-Quality Pseudo Data. To obtain image pairs of the same person wearing different garments, we use a mask-based try-on method for garment content replacement. However, mask-based methods inherently suffer from original content loss due to the mask. Therefore, we minimize the mask region to reduce the loss, as shown in Figure 1. Figure 2 and Figure 3 illustrate the pseudo data used during training from the VITON-HD [1] and DressCode[6] datasets, respectively. After optimizing the masks, more information unrelated to the try-on process is preserved, reducing the negative impact of mask-based methods on the quality of pseudo data. The coarse mask M_{coa} is generated using the same method as the masks in IDM-VTON [2].

Comparison of pseudo-data generation models. The generation of pseudo data is independent of IDM-VTON. We used CAT-DM and StableVITON to generate pseudo data, as shown in Figure 4-Left. The results of training with these types of pseudo data are presented in Table 1, where replacing the pseudo data generation model does not have a significant impact on the final model.

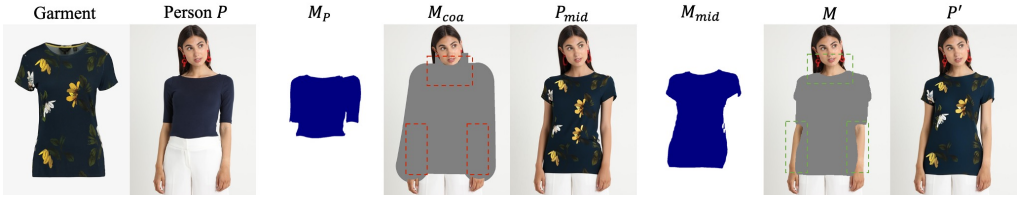

Figure 1: The mask M only covers the region where garment content changes, preserving more details of accessories and skin.

Testsets	**VITON-HD**							**StreetVTON**		**WildVTON**	
Model	**LPIPS** ↓	**SSIM** ↑	**PSNR** ↑	**FID$_p$** ↓	**KID$_p$** ↓	**FID$_u$** ↓	**KID$_u$** ↓	**FID$_u$** ↓	**KID$_u$** ↓	**FID$_u$** ↓	**KID$_u$** ↓
IDM-VTON+Wild Aug	0.1217	0.8530	21.08	8.489	2.537	9.271	1.256	23.71	6.193	39.31	7.794
CAT-DM-Pseudo	0.1091	0.8595	21.67	7.223	1.503	9.256	1.104	21.68	5.081	34.23	5.413
StableVITON-Pseudo	0.1085	0.8610	21.83	6.947	1.417	9.113	1.026	20.93	4.855	33.11	4.901
w/o Back-Aug	0.1074	0.8626	21.82	6.655	1.101	8.869	0.793	23.35	5.892	39.80	7.711

Table 1: Additional ablation experiments.

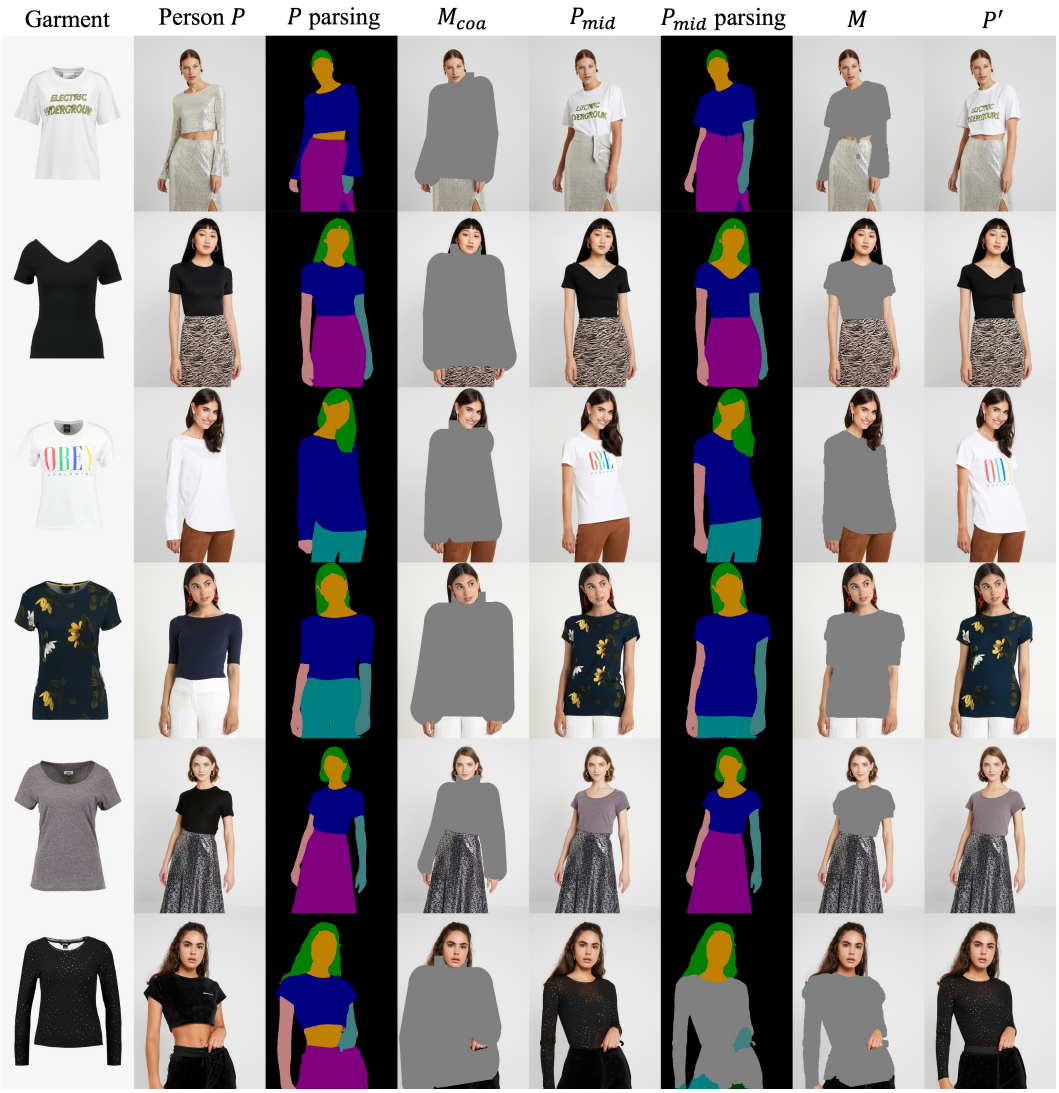

Figure 2: High-quality pseudo data generation in VITON-HD.

Figure 3: High-quality pseudo data generation in DressCode.

1.2 In-the-Wild Augmentation Details

In this section, we detail the implementation of wild data augmentation. In-the-Wild data augmentation utilizes 50 foreground images with a resolution of 1024×768, as shown in Figure 6, Figure 7, the corresponding prompts are shown in Table 3. And 40 background images with a resolution of 1024×1024, as shown in Figure 8, Figure 9, the corresponding prompts are shown in Table 4. We independently add foregrounds and backgrounds to the person images. We add a background with a probability of p=0.5, where the placement of the person within the background image is chosen randomly. A foreground is applied with a probability of p=0.7, with the foreground image being randomly scaled (0.25 to 0.6 times) and randomly rotated (-45 to 45 degrees) before being overlaid on the person image. Based on in-the-wild data augmentation, we applied the same spatial data augmentation as StableVITON [5]. As shown in Figure 5, we provide visualization results of the training samples used during the training process.

Explaination of In-the-wild Setting. Current mask-based approaches exhibit a strong dependency on mask precision, which leads to substantial performance deterioration in real-world scenarios where accurate mask acquisition is challenging. Therefore, even when trained on data with complex backgrounds, IDM-VTON fails to address the in-the-wild try-on problem. To demonstrate this, we re-trained IDM-VTON under wild-aug conditions. The results, shown in Table 1, indicate no significant change in IDM-VTON performance. Additionally, even without the complex background, BooW-VTON still outperforms the existing method, as shown in Table 2.

Modules	$\text{FID}_{\text{Street}} \downarrow$	$\text{KID}_{\text{Street}} \downarrow$	$\text{FID}_{\text{Wild}} \downarrow$	$\text{KID}_{\text{Wild}} \downarrow$
IDM-VTON	20.50	6.155	32.72	7.908
Ours	**19.47**	**5.513**	**31.75**	**7.386**

Table 2: Comparison results after replacing non-human regions in the in-the-wild test with white background.

Ablation of Background Augmentation. Back-Aug helps preserve complex backgrounds, as shown in Table 1 and Figure 4-Right. Moreover, “inpainting blank areas of the person image" refers to generating clean background images and only occurs during the background creation.

Figure 4: **Left:** Pseudo data of different generation model. **Right:** Without background augmentation, it leads to the loss of high-frequency details in complex backgrounds.

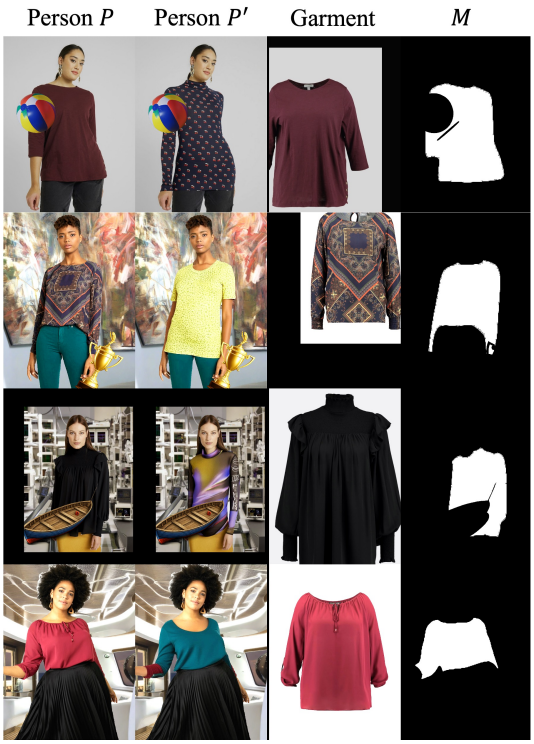

Figure 5: Data augmentation of training samples during the training process.

Figure 6: Foreground images used for in-the-wild data augmentation (Part 1).

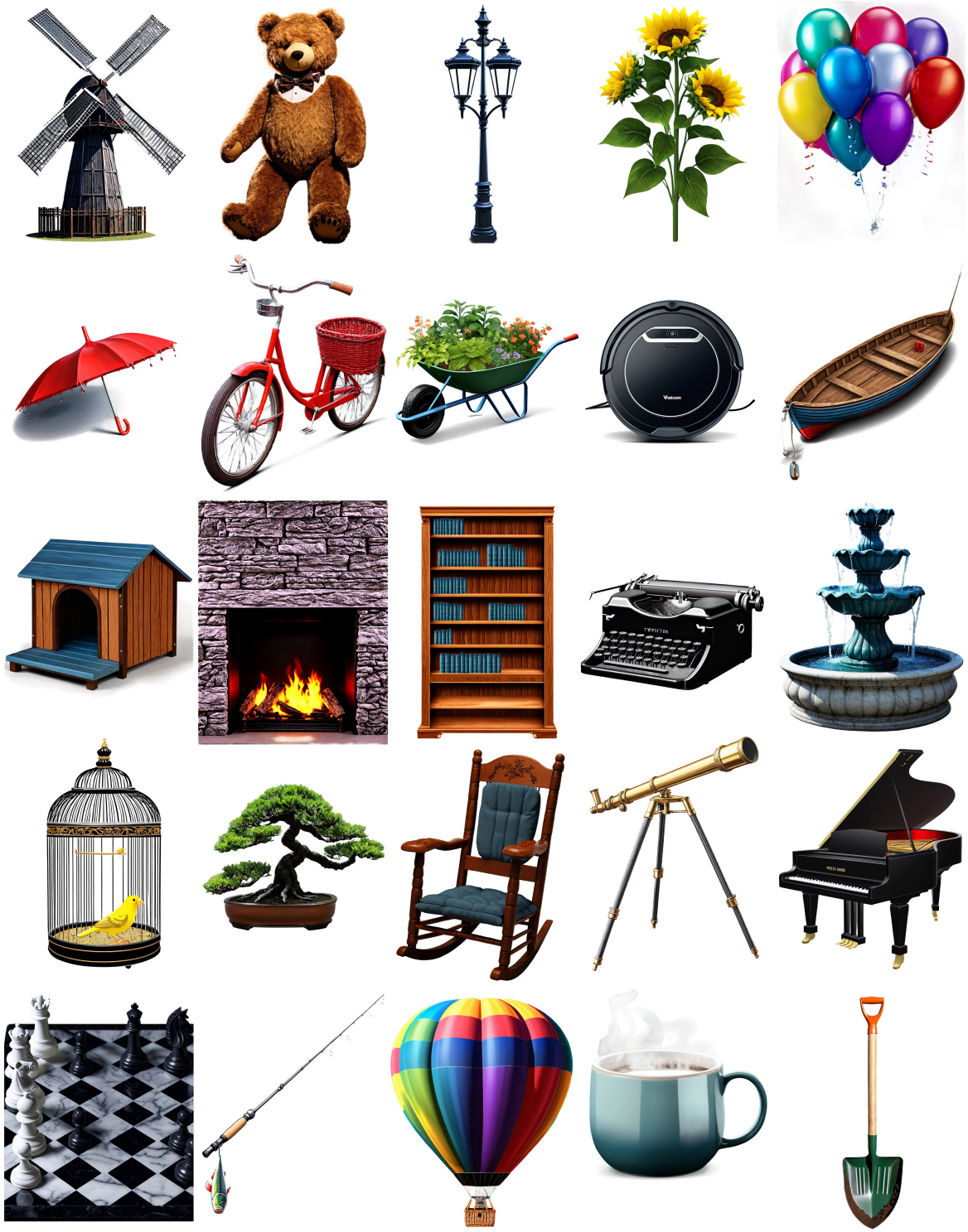

Figure 7: Foreground images used for in-the-wild data augmentation (Part 2).

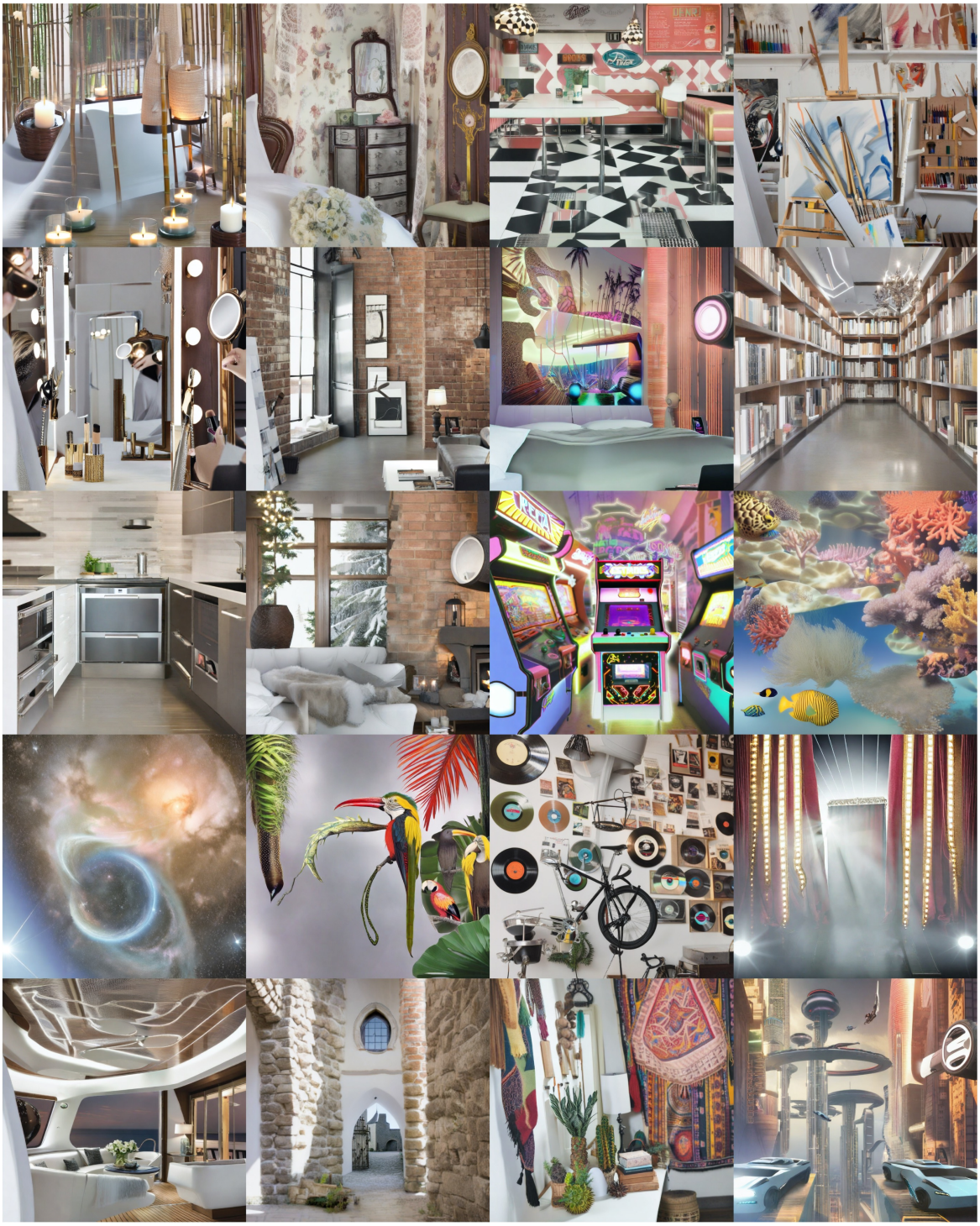

Figure 8: Background images used for in-the-wild data augmentation (Part 1).

Figure 9: Background images used for in-the-wild data augmentation (Part 2).

Table 3: Prompts for foreground images generation.

Category	Prompt
Foreground	Colorful balloons, helium-filled, shiny and reflective
	Sunflower bouquet, vibrant yellow, lush green stems
	Vintage bicycle, red paint, woven wicker basket
	Oversized teddy bear, brown fur, wearing a bow tie
	Classic street lamp, wrought iron, glowing light
	Red umbrella, open, raindrops dripping off edges
	Christmas wreath, adorned with pinecones and red ribbon
	Large suitcase, leather, covered in travel stickers
	Garden gnome, ceramic, holding a small lantern
	Floating lantern, paper, softly glowing candle inside
	Sparkling wine glass, crystal clear, filled with champagne
	Golden retriever puppy, playful, wearing a blue collar
	Picnic basket, woven, filled with fruits and bread
	Colorful kite, flying high, with long tail
	Beach ball, multi-colored, bouncing in the air
	Butterfly, vibrant wings, resting on a flower
	Basketball, orange, spinning on a finger
	Snow globe, intricate design, with a winter scene inside
	Vintage camera, black and silver, with a leather strap
	Book stack, old and new, with bookmarks sticking out
	Guitar, wooden, with intricate inlays
	Soccer ball, black and white, rolling on grass
	Ice cream cone, double scoop, with sprinkles
	Laptop, open, showing a bright screen
	Microphone with stand, black, rich details
	Cup of coffee, steaming, with latte art
	Bouquet of roses, red and white, with green leaves
	Candle, lit, with a gentle flame
	Championship trophy, gold, rich details
	Basket of apples, red and green, freshly picked
	Antique clock, wooden frame, intricate carvings
	Chessboard, marble, set up for a game
	Fishing rod, graphite, with a lure attached
	Hot air balloon, colorful, ascending into the sky
	Coffee mug, ceramic, with steam rising
	Garden shovel, metal, with soil on it
	Piano, grand, with open lid
	Telescope, brass, pointed towards the stars
	Rocking chair, wooden, with cushions
	Bonsai tree, miniature, with detailed branches
	Birdcage, ornate, with a singing canary
	Water fountain, marble, with flowing water
	Typewriter, vintage, with paper inserted
	Bookcase, wooden, filled with books
	Fireplace, stone, with roaring fire
	Dog house, wooden, with a nameplate
	Fishing boat, wooden, anchored near the shore
	Robot vacuum, sleek, cleaning the floor
	Garden wheelbarrow, metal, filled with plants
	Windmill, old-fashioned, turning in the wind

Table 4: Prompts for background images generation.

Category	Prompt
Background	{ } in a china garden
	{ } in a snowy winter landscape with pine trees
	{ } in a bustling urban street scene
	{ } in a garden, lush greenery
	{ } in front of a modern city skyline
	{ } on a sandy beach with ocean waves in the background
	{ } in a cozy cafe setting with coffee cups and books
	{ } in an elegant, upscale interior with chandeliers
	{ } in a rustic countryside setting with fields and barns
	{ } in a futuristic environment with metallic structures
	{ } in a vintage-inspired room with old furniture and antiques
	{ } in a snowy winter landscape with pine trees
	{ } in a high-tech laboratory setting with monitors and equipment
	{ } in a chic, modern art gallery
	{ } in a traditional Japanese garden with pagodas and ponds
	{ } in a busy airport terminal with travelers and luggage
	{ } in an industrial warehouse with machinery and crates
	{ } in a mystical forest setting with fog and ancient trees
	{ } in a sports stadium during a match
	{ } in a dramatic desert landscape with sand dunes and sunset
	{ } in a futuristic cityscape with flying cars
	{ } in a bohemian-style room with colorful tapestries
	{ } in a medieval castle courtyard with stone walls
	{ } in a luxury yacht setting with ocean views
	{ } in a classical theater with velvet curtains and stage lights
	{ } in a hipster cafe filled with vintage bicycles and records
	{ } in a tropical rainforest with exotic plants and wildlife
	{ } in a celestial-themed setting with stars and galaxies
	{ } in an underwater scene with coral reefs and tropical fish
	{ } in a 1980s retro arcade with neon lights and arcade games
	{ } in a cozy living room with a fireplace and soft blankets
	{ } in a sleek modern kitchen with stainless steel appliances
	{ } in a sophisticated library surrounded by shelves of books
	{ } in a futuristic bedroom with high-tech gadgets and mood lighting
	{ } in a chic urban loft with exposed brick walls and industrial decor
	{ } in a glamorous dressing room with mirrors and vanity lights
	{ } in a minimalist art studio with canvases and paintbrushes
	{ } in a retro diner with vinyl booths and checkerboard floors
	{ } in a vintage-themed boudoir with lace curtains and antique furniture
	{ } in a spa setting with candles, bamboo, and relaxation areas

2 Ablation Study

2.1 Additional Ablation Study Results

Figure 10: **Left**: Ablation study on the VITON-HD test set shows that H.Q. pseudo data enhances the retention of skin details and accessories. **Right**: Ablation study on the WildVTON test set demonstrates that in-the-wild data augmentation significantly improves the retention of both foreground and background elements.

2.2 Study on the Effectiveness of Attention Layers

Figure 11 presents an ablation study of attention layers in BooW-VTON. We selected steps 1, 5, 10, 15, 20, 25, and 30 from the DDIM sampler for analysis. We examined the cross-attention and self-attention layers from attention blocks 1, 25, 35, and 70 for each denoising step. As shown in Figure 11, the model focuses on the distribution of clothing in the initial denoising stages, while focusing on the details of the garment in later stages. In the network, the attention layers near the middle stages exhibit greater activity.

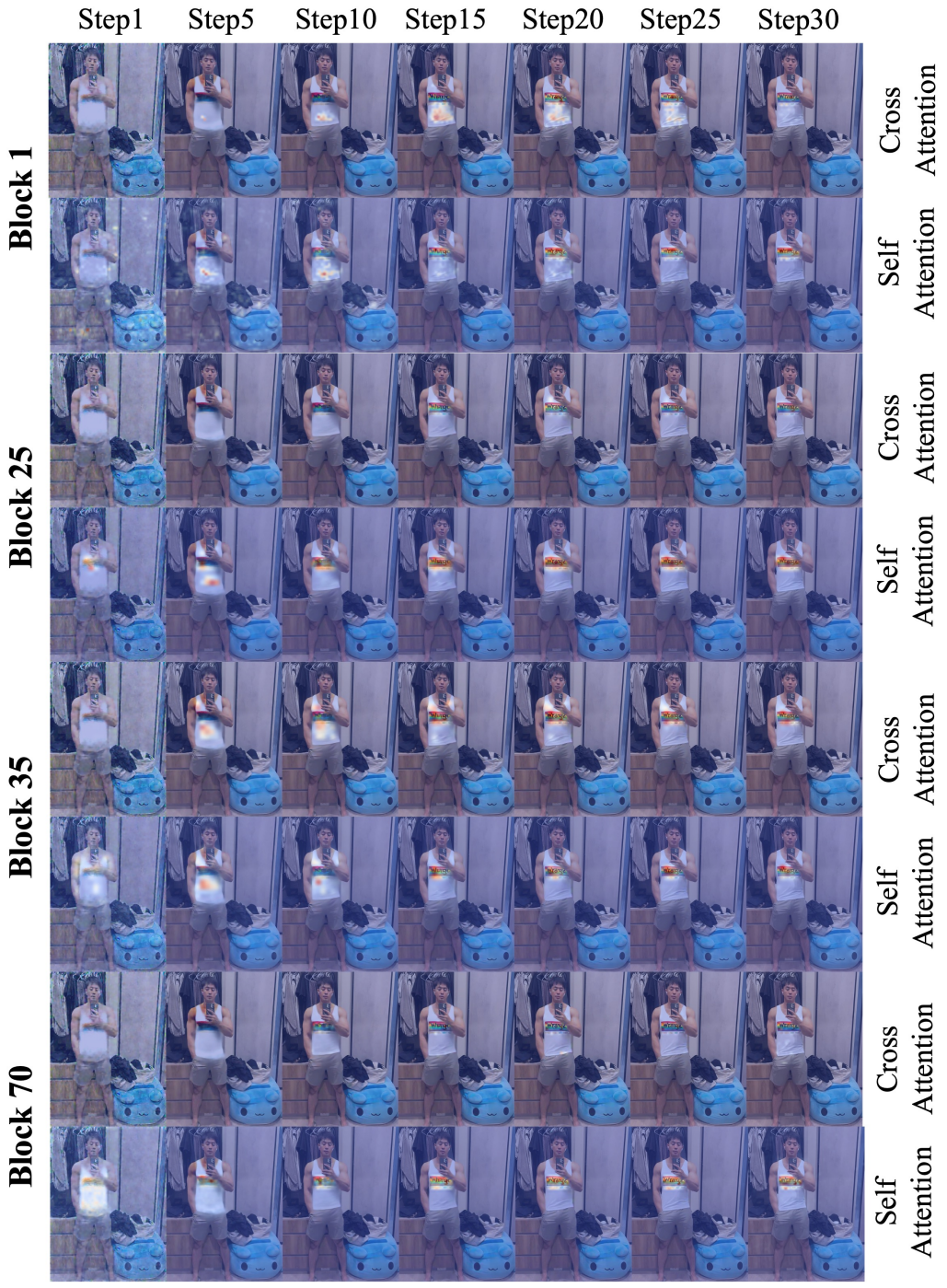

Figure 11: Visualization of attention heatmaps across different denoising stages and attention layers.

2.3 Attention Visualization in Ablation Study

We conducted ablation studies on the modules by visualizing the attention maps. As shown in Figure 12, directly applying a mask-free approach fails to guide the model to focus on the correct try-on regions during the wild try-on, resulting in the scattering of attention across the entire image. After incorporating in-the-wild data augmentation, the model learns to retain content in non-try-on regions, shifting dispersed attention toward the try-on areas. With the application of the try-on localization loss, the attention maps are explicitly regularized, enabling the model to identify the try-on regions more accurately.

Figure 12: Cross-attention maps and self-attention maps in ablation experiments.

3 Dataset Information

This section provides detailed information and image examples of the datasets used in the main text. As shown in 13, the VITON-HD [1] dataset consists of samples of women's spring and summer tops, with the training set containing 11,647 pairs and the test set containing 2,032 pairs. The DressCode [6] dataset is divided into three garment types: tops, bottoms, and dresses. It includes 29,478 dress samples, 15,363 top samples, and 8,951 bottom samples, with 1,800 pairs from each category extracted for the test set, as shown in 16. StreetVTON [3] is a subset selected from DeepFashion2 [4] for in-the-wild try-on, as shown in 14, containing 2,089 complex person images. WildVTON consists of 1224 lifestyle portrait images with complex foregrounds and backgrounds that we collected from the internet, as shown in 15. In addition, we validated the in-shop try-on performance on the VITON-HD and all three garment type DressCode test sets.

Figure 13: Samples of VITON-HD.

Figure 14: Samples of StreetVTON.

Figure 15: Samples of WildVTON.

Figure 16: Samples of DressCode.

4 More Visual Results

4.1 VITON-HD

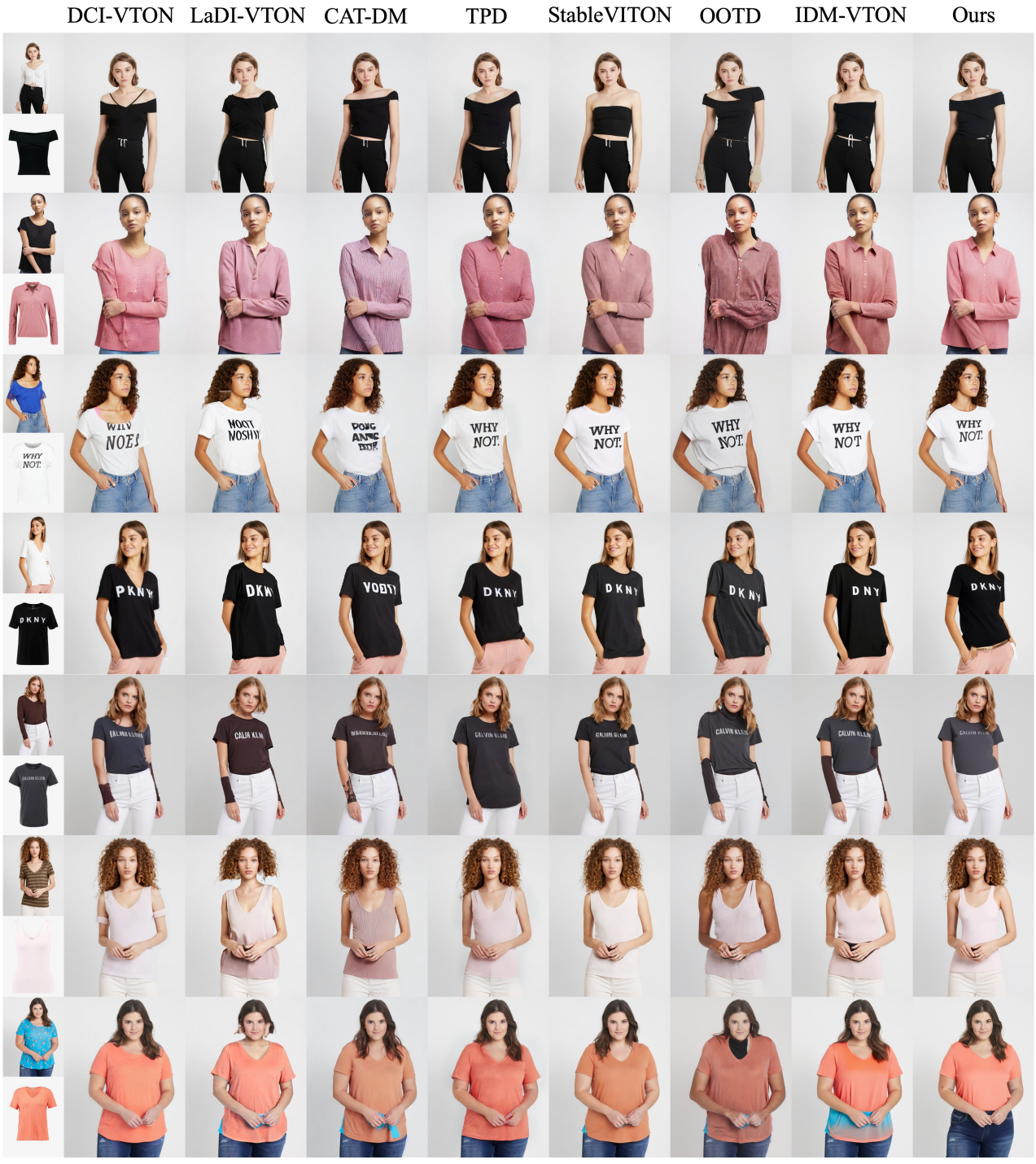

Figure 17: Visualization of comparison results on the VITON-HD test set.

Figure 18: Additional try-on results on the VITON-HD test set.

4.2 DressCode

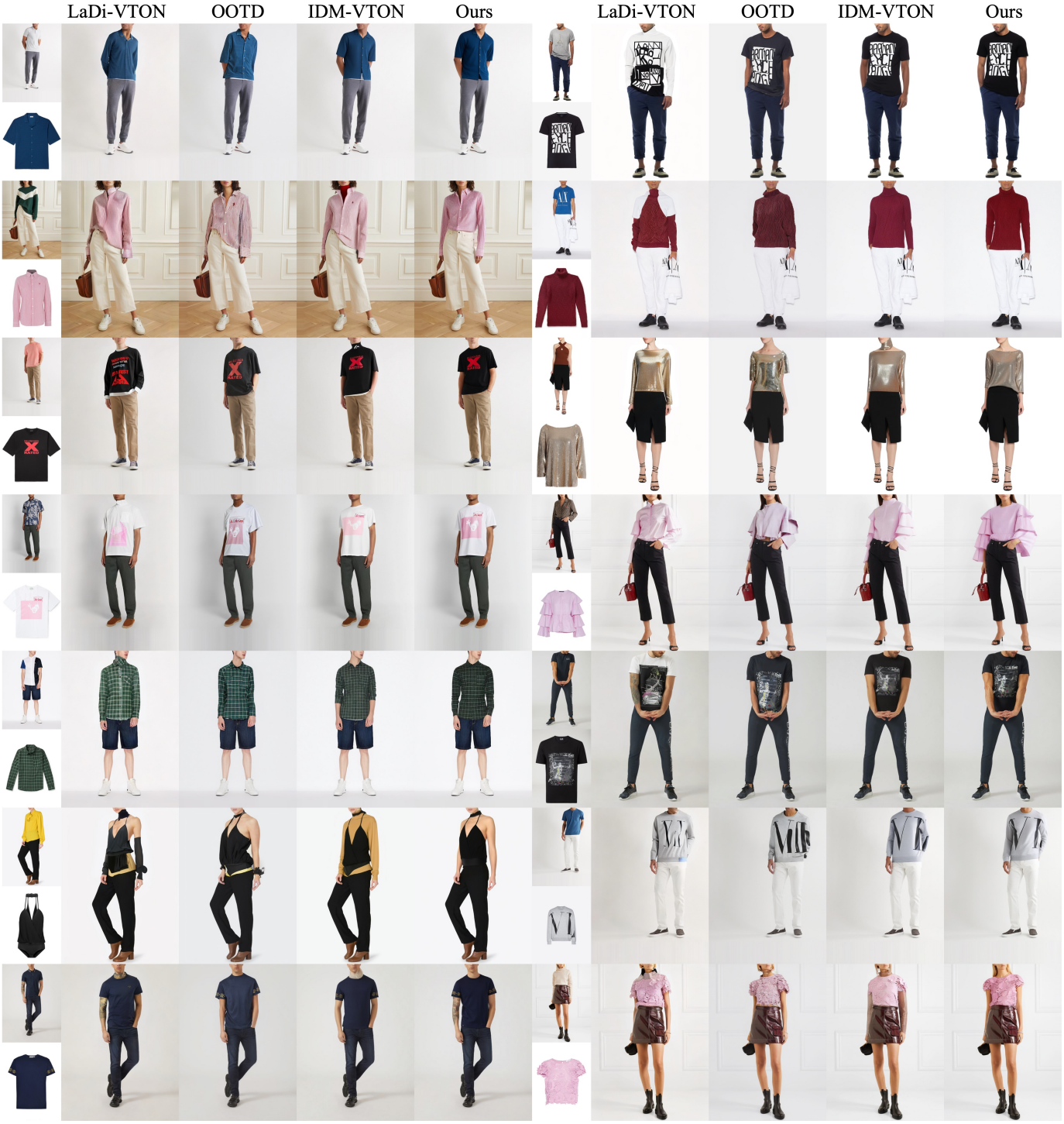

Figure 19: Visualization of comparison results on the DressCode Upper test set.

Figure 20: Visualization of comparison results on the DressCode Lower test set.

Figure 21: Visualization of comparison results on the DressCode Dresses test set.

Figure 22: Cross-Garment type try-on results on the DressCode test set (Part 1).

Figure 23: Cross-Garment type try-on results on the DressCode test set (Part 2).

4.3 StreetVTON

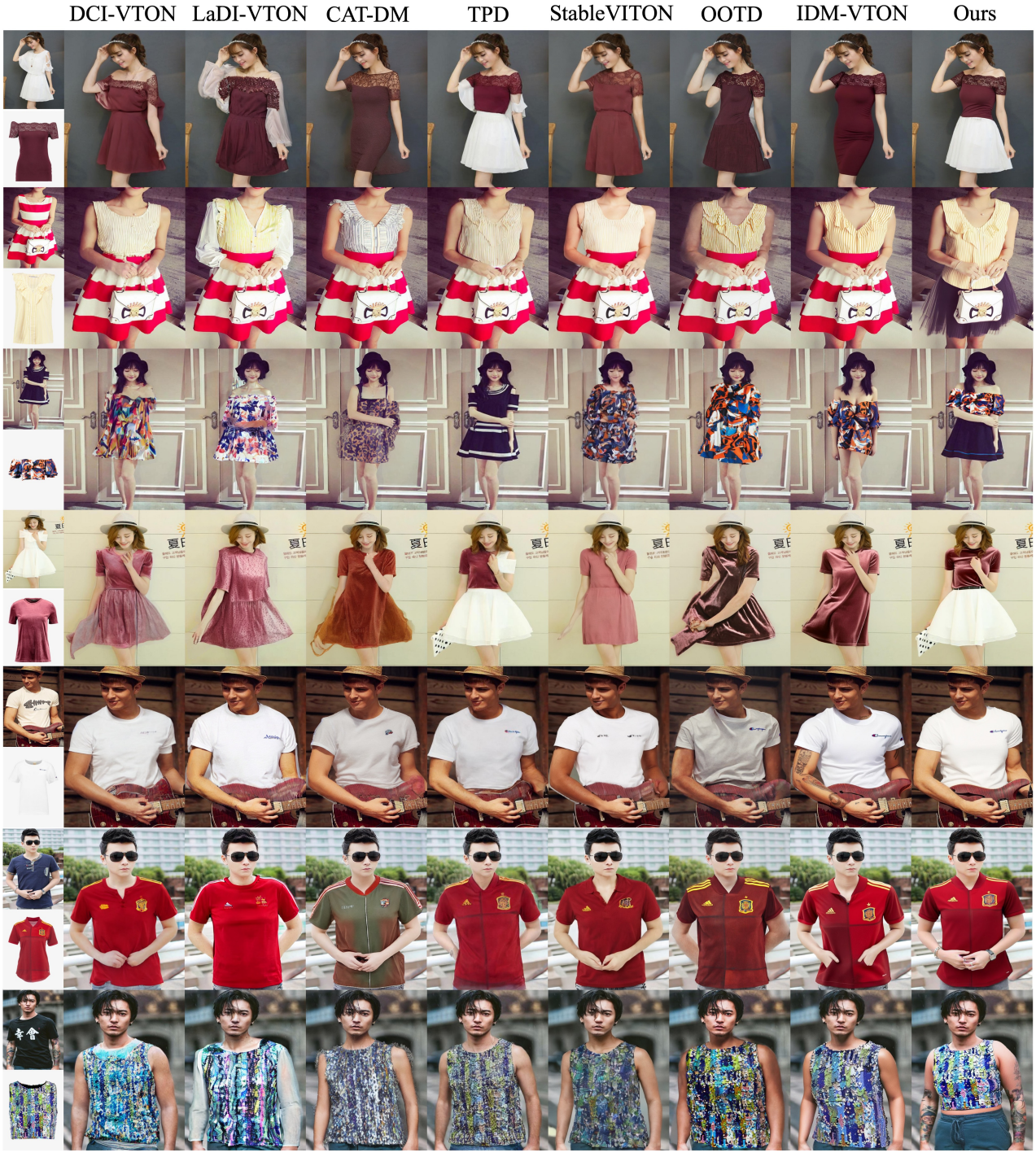

Figure 24: Visualization of comparison results on the StreetVTON test set.

4.4 WildVTON

Figure 25: Visualization of comparison results on the WildVTON test set.

Figure 26: Cross-Garment type try-on results on the WildVTON test set.

4.5 High-quality In-the-wild Try-on

BooW-VTON achieves highly realistic and natural in-the-wild try-on results. This realism stems from the preservation of lighting conditions and spatial structures. As shown clearly in Figure 27: In the first row, our try-on results not only retain the details of the subject's skin but also faithfully reproduce the shading on the arm. In the second row, while preserving details across various types of garment, BooW-VTON maintains the influence of the top light source on the entire image. In the third row, BooW-VTON accounts for the model's orientation and leg movement, deforming the clothing accordingly to produce realistic and natural try-on results.

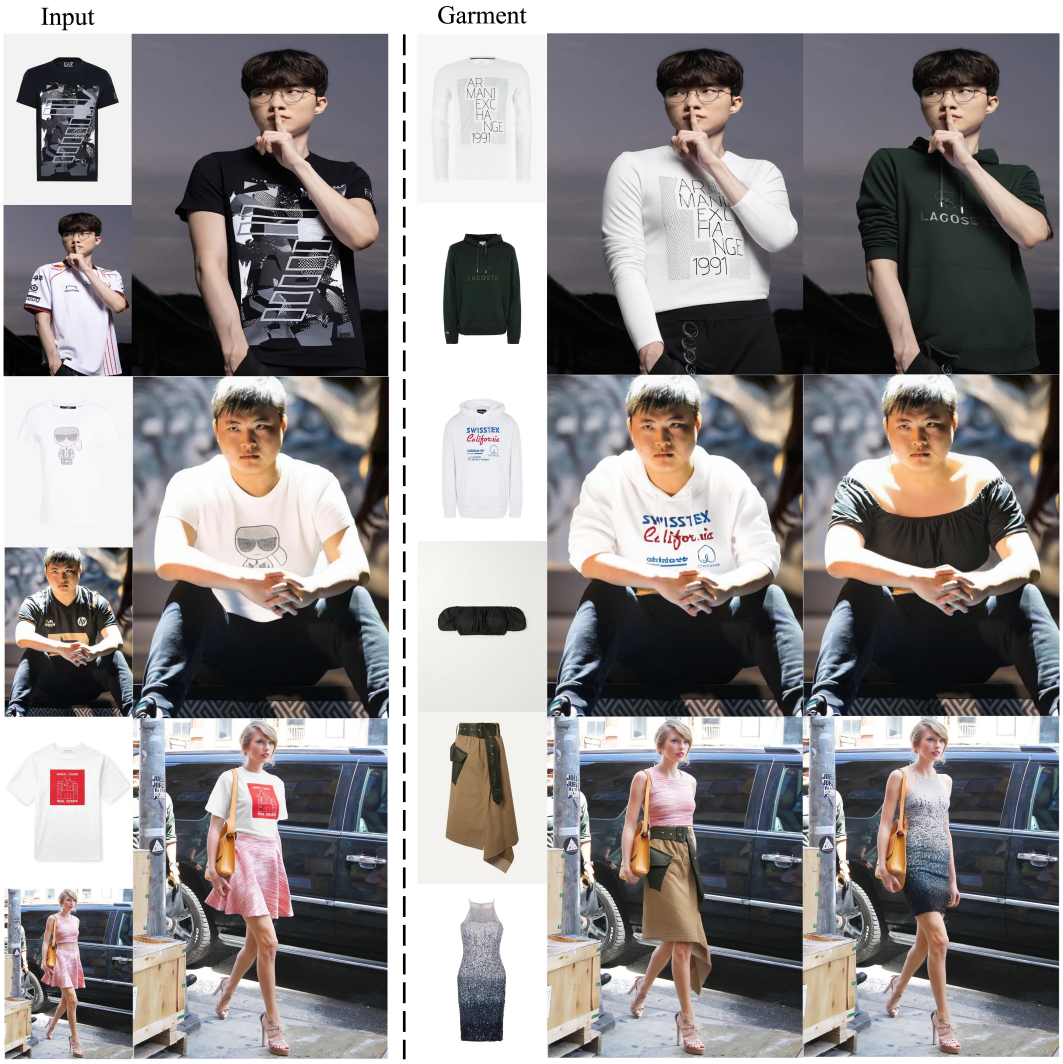

Figure 27: High-quality in-the-wild try-on results with lighting features and spatial structures.

5 Real-time Performance

We perform inference in bfloat16 mode on an Nvidia A100, with the inference time and memory usage as follows:

Modules	AutoEncoder	TextEncoder	IP-Adapter	Garment-Unet	Try-On-Unet	Total
Inference Time(ms)	386	452	34.1	79.8*30	100.3*30	6241
GPU Memory(MB)	170	1628	1314	5144	5984	14240

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.

[2] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for authentic virtual try-on in the wild. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXVI*, volume 15144 of *Lecture Notes in Computer Science*, pages 206–235. Springer, 2024.

[3] Aiyu Cui, Jay Mahajan, Viraj Shah, Preeti Gomathinayagam, and Svetlana Lazebnik. Street tryon: Learning in-the-wild virtual try-on from unpaired person images. *CoRR*, abs/2311.16094, 2023.

[4] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5337–5345. Computer Vision Foundation / IEEE, 2019.

[5] Jeongho Kim, Gyojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. *CoRR*, abs/2312.01725, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2231–2235, 2022.