High-Fidelity Virtual Try-On beyond Paired Data Scarcity via Diffusion-based Cycle-Consistent Learning

Supplementary Material

6. Implementation Details

6.1. Training Data

For training, we adopt a three-stage pipeline:

- **Stage (a): Paired pretraining.** We train the Unified Diffusion Transformer (UDiT) on the official training splits of VITON-HD (11647 images) and DressCode (48392 images), following standard data partitioning.
- **Stage (b): Wild data filtering.** We collect in-the-wild person images from publicly available fashion websites and retain 1M images after initial VLM-based filtering. Using the UDiT trained in Stage (a), we perform try-off extraction followed by the multi-criteria filtering operation. After filtering, approximately 70% (700K) of the samples are retained as high-quality pseudo-labeled garment-person pairs.
- **Stage (c): Cycle-consistent learning.** We fine-tune the UDiT on a hybrid training set that combines the original paired training data (VITON-HD + DressCode, 60K samples) and the filtered wild data (700K samples). This mixed supervision enables the model to generalize across both controlled and unconstrained settings.

6.2. Implementation Setting

Our method is based on the dev version of FLUX.1.Fill model. We train stage (a) with batch-size of 128 for 5K steps, and stage (c) with batch-size of 128 for 30K steps. All the experiments are conducted on 32 NVIDIA H20 GPUs. We employ the AdamW optimizer with learning rate of $1.5e^{-5}$ and weight decay of $1e^{-4}$ for training stages. The parameters of λ and β are set as 1.0 and 0.1, respectively. For inference, we adopt 35 denoising steps (10 for Stage 1, 25 for Stage 2) to generate the results at a resolution of 1024×768.

7. Prompts

We provide the prompt templates for different stages of CCVTON, which involves the input of UDiT, selection of in-the-wild person images, and the judgment of MCFO.

7.1. Prompt for UDiT

Tab. 3 presents the text prompt template used in both the training and inference phases of the UDiT. This unified prompt is employed for both try-on and try-off tasks, ensuring consistent semantic guidance across different stages of the pipeline. The placeholder {category} is dynamically replaced with the specific garment category (e.g., "Upper", "Lower", or "Whole") during processing.

7.2. Selection for In-the-Wild Person Images

Tab. 4 presents the prompt template used to filter in-the-wild person images via the Qwen2.5-VL-32B. This structured prompt instructs the model to act as a fashion director, categorizing each {PERSON IMAGE} based on four key criteria: *HumanPresence*, *Gender*, *BodyPartShown*, and *Watermark*. For each criterion, the model selects the most appropriate label from a predefined list, ensuring consistent and interpretable annotations. Specifically, we retain only those images that satisfy the following conditions: (1) *HumanPresence* = "Yes"; (2) *Gender* = "Male" or "Female"; (3) *HumanCount* = "Single"; (4) *BodyPartShown* = "Whole", "Upper", or "Lower"; and (5) *Watermark* = "No". Through this process, we obtain a dataset of 1 million in-the-wild person images.

7.3. Details of MCFO Filtering

Prompt for VLM judgement. After extracting garment images from 1 million in-the-wild person images using the pre-trained UDiT model (Stage (a)), we form person-garment image pairs <{PERSON IMAGE}, {GARMENT IMAGE}> and feed them into Qwen2.5-VL-32B for semantic consistency evaluation. The prompt in Tab. 5 guides the VLM to assess four key attributes: *PatternDifference*, *PatternRank*, *ColorDifference*, and *OtherObject*. We retain only those pairs that satisfy the following stringent criteria: (1) *PatternDifference* = "No"; (2) *PatternRank* = "No"; (3) *ColorDifference* = "No"; and (4) *OtherObject* = "No".

To further ensure visual similarity, we apply a ViT-based feature similarity threshold of 0.95 between the person and extracted garment images. By combining the VLM judgment with this metric, we filter out inconsistent or low-quality samples, ultimately retaining approximately 70% of the initial dataset. This two-stage filtering strategy effectively constructs a high-fidelity pseudo-paired dataset suitable for robust virtual try-on training.

8. More Experiments

8.1. Ablation Results

As shown in Fig. 9, we present qualitative results of the ablation study. The model variants are: **Exp.0 (BASE)**: Trained on standard paired datasets using a cloth-agnostic mask. **Exp.1**: Uses the cloth-bbox mask during training

Input of UDiT

The pair of images highlights a {category} clothe and its styling on a model, high resolution, 4K, 8K;
[CLOTHE IMAGE] Detailed product shot of a {category} clothe;
[MODEL IMAGE] The same {category} clothe is worn by a model in a lifestyle setting.

Table 3. Prompt template for the training and inference of UDiT.

MCFO / Selection for In-the-Wild Person Images

Input:
IMAGE:
{PERSON IMAGE}
PROMPT:
You are a fashion director. Your task is to categorize the image of individual models based on following detailed criteria. For each criterion, choose the most appropriate label from the provided list.

HumanPresence: ["Yes", "No", "Uncertain"]
"Yes": The image clearly contains a human (model).
"No": The image does not contain a human.
"Uncertain": It is unclear whether the image contains a human.

Gender: ["Male", "Female", "Uncertain"]
"Male": Clearly male.
"Female": Clearly female.
"Uncertain": Cannot determine gender.

HumanCount: ["Single", "Multiple", "Uncertain", "NAN"]
"Single": Exactly one distinct human model is visible in the image.
"Multiple": Two or more distinct human models are present, including cases where the same model appears in multiple views.
"Uncertain": Human presence is ambiguous or the number of models cannot be reliably determined.
"NAN": No human model is present (e.g., only garments on hangers, product shots, or non-human subjects).

BodyPartShown: ["Whole", "Upper", "Lower", "Other"]
"Whole": The entire body is visible.
"Upper": The upper body (from head to waist) is visible.
"Lower": The lower body (from waist to feet) is visible.
"Other": Other body parts (e.g., face, hands) are shown.

Watermark: ["Yes", "No", "Uncertain"]
"Yes": The image has a visible watermark.
"No": The image does not have a watermark.
"Uncertain": It is unclear whether the image has a watermark.

Output only the **JSON** dictionary, with no additional text. Example format:
{"HumanPresence":"Yes", "Gender":"Female", "HumanCount": "Single", "BodyPartShown":"Whole", "Watermark":"No"}

Output: {answer}

Table 4. Prompt for filtering in-the-wild person images.

and applies our GAMG at inference. **Exp.2**: Introduces the UDiT, jointly optimizing try-on and try-off branches end-to-end on paired data. **Exp.3**: Extends Exp.2 by leveraging UDiT to extract garment representations from large-scale in-the-wild person images. These are filtered via MCFO to form high-quality pseudo-paired data for further fine-tuning. **Exp.4**: Trains UDiT via CCL, using a try-off and try-on reconstruction cycle on wild images. **Exp.5**: CCVTON. Specifically, Exp.0, trained with a cloth-agnostic mask, suffers from severe garment leakage and texture distortion. Exp.1 improves structural consistency by using a cloth-bbox mask and applying GAMG at inference, reducing leakage but still producing inaccurate leg coverage and mismatched proportions. Exp.2 and Exp.3 still exhibit inconsistencies in clothe style adaptation. Exp.4 employs CCL on in-the-wild data, achieving strong structural coherence and garment texture preservation, but the problem of hand collapse still exists. Finally, Exp.5 combines all

MCFO / Prompt for VLM judgement

Input:
IMAGE PAIR:
{PERSON IMAGE}
{GARMENT IMAGE}
PROMPT:
You are a fashion director. Your task is to compare the two garment images and provide a description using a JSON dictionary. The attributes to be judged are as follows:

PatternDifference: ["Yes", "No", "Uncertain"]
Determine if there is a difference in patterns or prints between the two images.
PatternRank: ["Yes", "No", "Uncertain"]
Determine if there is a difference in the arrangement of pattern elements between the two images.
ColorDifference: ["Yes", "No", "Uncertain"]
Determine if there is a color discrepancy between the garments in the two images.
OtherObject: ["Yes", "No", "Uncertain"]
Determine if there are any non-garment objects present in the second image.

Output only the **JSON** dictionary with the four keys listed above, strictly using labels from the provided options. Example format:
{"PatternDifference":"Uncertain", "PatternRank":"Uncertain", "ColorDifference":"Uncertain", "OtherObject":"No"}

Output: {answer}

Table 5. Prompt for VLM judgement based on person-garment pairs.

Figure 9. Ablation results on a in-the-wild person-garment data.

components and achieves the most visually plausible results: accurate garment fitting, preserved body structure,

and seamless integration with the original pose and background. These results demonstrate that each component contributes incrementally to high-fidelity virtual try-on in challenging real-world scenarios.

8.2. Parameter Sensitivity

We investigate the sensitivity of the corresponding weights in the loss function, i.e., β for perceptual regularization. The results on the VITON-HD dataset are shown in Fig. 10. We try several values for the weight parameter from $\{0.0, 0.1, 0.3, 0.5, 0.7, 1.0\}$ during cycle-consistent learning. It can be found that when $\beta = 0.1$, the model achieves the best performance. This suggests that although the perceptual regularization effectively guides the try-off branch, excessively strong supervision may impede model generalization, thereby ultimately compromising the quality of the generated output. Thus, we use it as the default loss weight during the training phase.

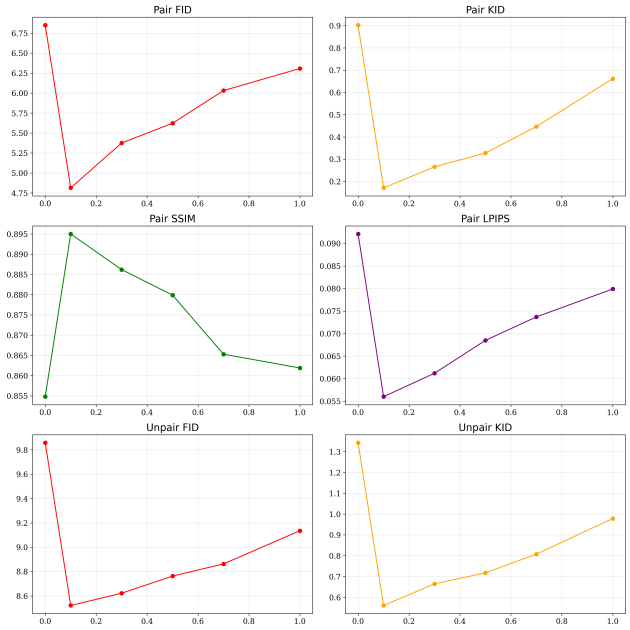

Figure 10. Sensitivity of parameter β on the VITON-HD dataset. For all metrics but SSIM, lower values indicate better performance.

	TryOffDiff	TryOffAnyOne	Any2AnyTryon	**Ours**
FID↓	28.25	25.20	13.37	13.29
KID↓	11.42	6.98	3.50	3.60

Table 6. Quantitative comparison for try-off on VITON-HD.

8.3. More Comparison with Baselines

Quantitative Results on DressCode. Tab. 7 shows the quantitative comparisons of CCVTON with other methods on the DressCode dataset. As shown in the table, CCVTON achieves the best score on all metrics in each category setting. In the Paired setting, our method consistently outperforms existing approaches by a significant margin. For instance, on DressCode-Upper, CCVTON obtains the lower FID and KID, demonstrating superior perceptual realism and distributional fidelity. This improvement is further validated by higher SSIM and lower LPIPS, indicating enhanced structural alignment and visual similarity to ground-truth garments. Similar gains are observed across the Lower and Dress categories, confirming the generalizability of our design. In the Unpaired setting, where no paired supervision is available during training, CCVTON still achieves state-of-the-art performance. Notably, it obtains the lowest FID and KID scores in all three subcategories, indicating that our CCL strategy effectively learns robust garment representations from unpaired wild data without relying on explicit pairing. The consistent superiority across both settings underscores the effectiveness of our cycle-consistent learning strategy. Moreover, the marginal improvement from "Our (w/o GAMG)" to "Ours" highlights the critical role of GAMG in refining the mask area, reducing texture leakage, and preserving body structure.

More Comparison on In-the-Wild Data. As shown in Fig. 13 and Fig. 14, we present more qualitative comparisons in real-world scenarios, which demonstrate the superiority of our CCVTON.

8.4. Experimental Results for Try-Off

Quantitative Results. Tab.6 presents the quantitative comparison of our CCVTON against state-of-the-art try-off methods on the VITON-HD dataset. These results demonstrate that our CCVTON framework effectively generates high-fidelity try-off outputs, achieving competitive or superior performance against existing methods while maintaining strong generalization to real-world scenarios.

Qualitative Results. Fig. 11 presents the try-off results of CCVTON on in-the-wild person images, demonstrating its ability to accurately extract garments under diverse real-world conditions. The three columns correspond to upper-body, lower-body, and full-dress categories, respectively. As shown, our model successfully identifies and isolates target garment regions across a wide range of clothing types, while preserving fine details like stitching, logos, and fabric patterns. Notably, even in challenging cases involving occlusions (e.g., hands covering parts of the garment), non-frontal poses, or irregular lighting, the extracted garments remain visually coherent and structurally faithful. For example, the ruffled sleeves and V-neckline of the blue dress are preserved with high fidelity, and the intricate embroidery on the qipao is clearly retained. These results validate the robustness and generalization capability of our UDiT-based try-off module, which leverages both structural disentanglement and self-supervised learning to produce high-

Method	DressCode-Upper						DressCode-Lower						DressCode-Dress					
	Paired				Unpaired		Paired				Unpaired		Paired				Unpaired	
	FID↓	KID↓	SSIM↑	LPIPS↓	FID↓	KID↓	FID↓	KID↓	SSIM↑	LPIPS↓	FID↓	KID↓	FID↓	KID↓	SSIM↑	LPIPS↓	FID↓	KID↓
GP-VTON	16.282	7.265	0.910	0.080	20.162	9.389	15.400	4.726	0.898	0.085	20.769	6.703	18.491	9.474	0.823	0.143	24.325	14.196
IDM-VTON	10.435	1.401	0.909	0.061	12.961	1.520	13.699	5.112	0.893	0.067	18.277	6.626	14.610	2.913	0.843	0.102	17.328	5.873
OOTDiffusion	9.894	1.416	0.907	0.065	14.581	2.565	10.580	1.507	0.890	0.061	15.621	5.258	11.299	2.029	0.846	0.097	16.960	5.774
Any2anyTryon	11.936	1.945	0.877	0.085	15.673	4.510	14.020	5.716	0.833	0.988	20.399	7.790	14.788	3.135	0.801	0.114	21.563	10.797
CatVTON	11.107	1.632	0.898	0.078	12.172	3.058	11.900	3.969	0.895	0.064	17.430	5.996	13.858	3.908	0.837	0.117	18.380	7.266
PromptDresser	9.659	1.325	0.907	0.068	11.508	0.713	10.710	1.302	0.876	0.084	19.342	5.659	11.964	2.884	0.840	0.110	20.493	10.091
Leffa	8.057	0.655	0.909	0.047	11.388	0.618	10.263	1.557	0.879	0.067	14.300	3.133	13.902	3.810	0.842	0.106	17.920	6.582
Our (w/o GAMG)	6.873	0.550	0.927	0.036	11.434	0.652	9.133	1.280	0.900	0.058	14.560	3.232	9.534	1.561	0.857	0.078	13.268	4.687
Ours	**5.890**	**0.146**	**0.941**	**0.029**	**10.993**	**0.797**	**7.934**	**1.174**	**0.922**	**0.039**	**13.995**	**2.912**	**7.841**	**0.924**	**0.870**	**0.059**	**14.046**	**3.371**

Table 7. Quantitative comparison with other methods on the each category of DressCode dataset. We report both paired and unpaired evaluation results. The best and second-best results are highlighted in **bold** and underlined, respectively.

quality pseudo-garment representations from uncontrolled wild data.

8.5. Try-On Generalization

Fig. 12 presents a comprehensive set of try-on results generated by CCVTON across diverse garments and individuals, demonstrating its strong generalization capability in real-world scenarios. The first row shows the input garments, while subsequent rows display the same garments transferred to different person images with varying body shapes, poses, and backgrounds.

As shown, our method successfully transfers a wide range of clothing styles to models in varied poses, maintaining accurate silhouette alignment and fabric deformation. Notably, even for garments with complex patterns or intricate textures, the model preserves details and avoids common artifacts, such as texture blurring, misalignment, or unnatural stretching. Moreover, the well generation results in different individuals show that CCVTON effectively handles cross-category and cross-style try-on tasks. These results highlight the effectiveness of our CCVTON, enabling high-quality virtual try-on across diverse garments and human poses.

9. User Study

This section provides the details of user study. Specifically, participants were shown a reference garment and person image, along with four generated try-on results from our method and other three state-of-the-art baselines. The survey randomly presented 50 sets of generated results to each participant. For each set of results displayed in the survey, we ensured that their order was randomly shuffled to prevent bias. A screenshot of the evaluation interface is show in Fig. 15, which includes visual examples and five questions:

1. *Which result best preserves the **garment texture**?*
   — Consider the fidelity of fabric patterns, color consistency, and surface details.
2. *Which result best maintains the **garment shape**?*
   — Assess how accurately the cut, drape, and structural form (e.g., sleeve length, waistline, fit) of the garment are rendered on the person.
3. *Which result best retains **person details**?*
   — Focus on the preservation of body pose, skin tone, facial features, and non-target clothing regions.
4. *Which result appears the most **photo-realistic**?*
   — Judge overall visual plausibility, including lighting consistency, shadow coherence, and naturalness of the composite image.
5. *Which result is your **overall preferred choice** based on all aspects above?*

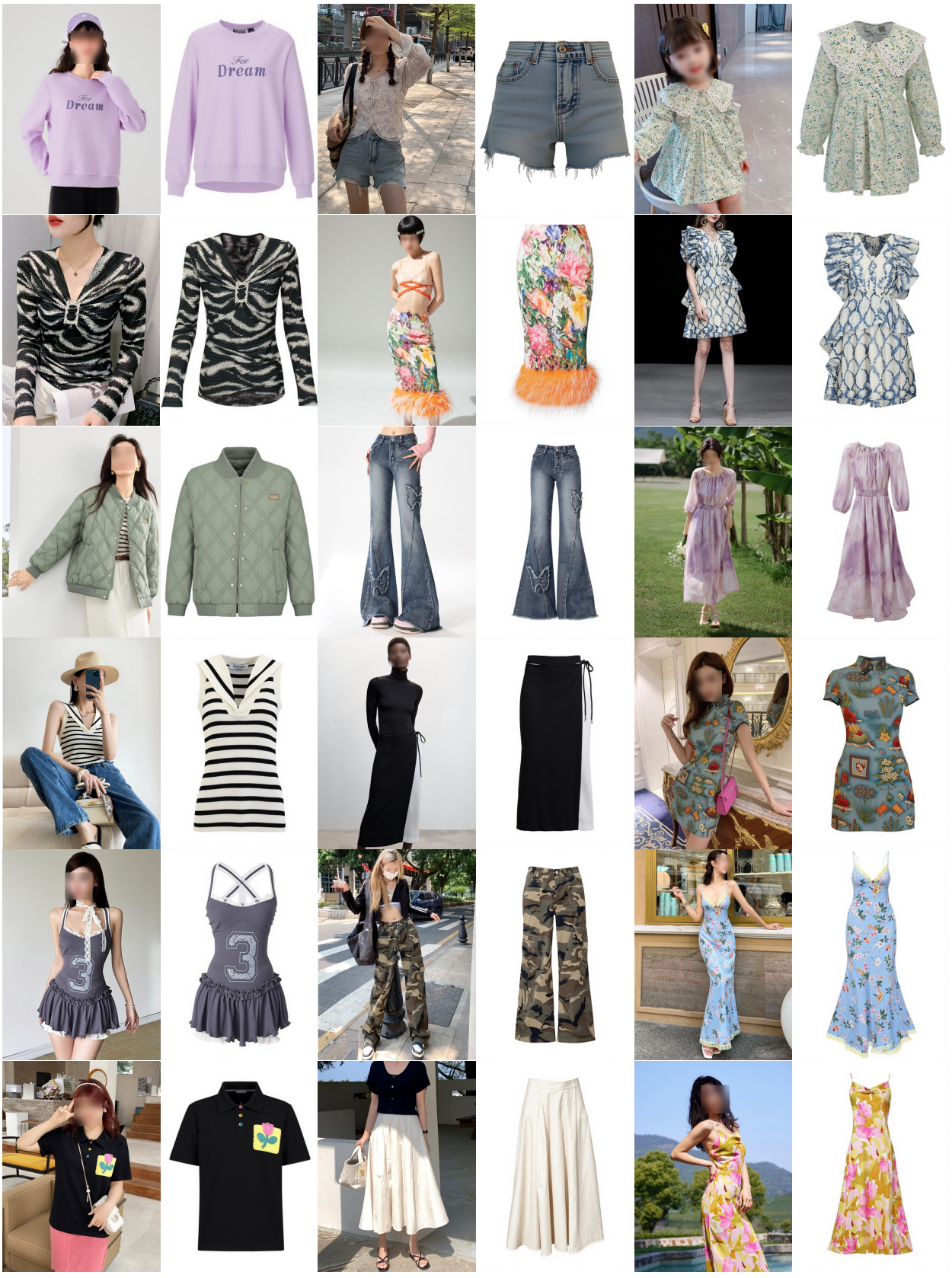

Figure 11. Try-off results on the wild data. Please zoom in for more details.

Figure 12. More try-on results of CCVTON across different garments and individuals. Please zoom in for more details.

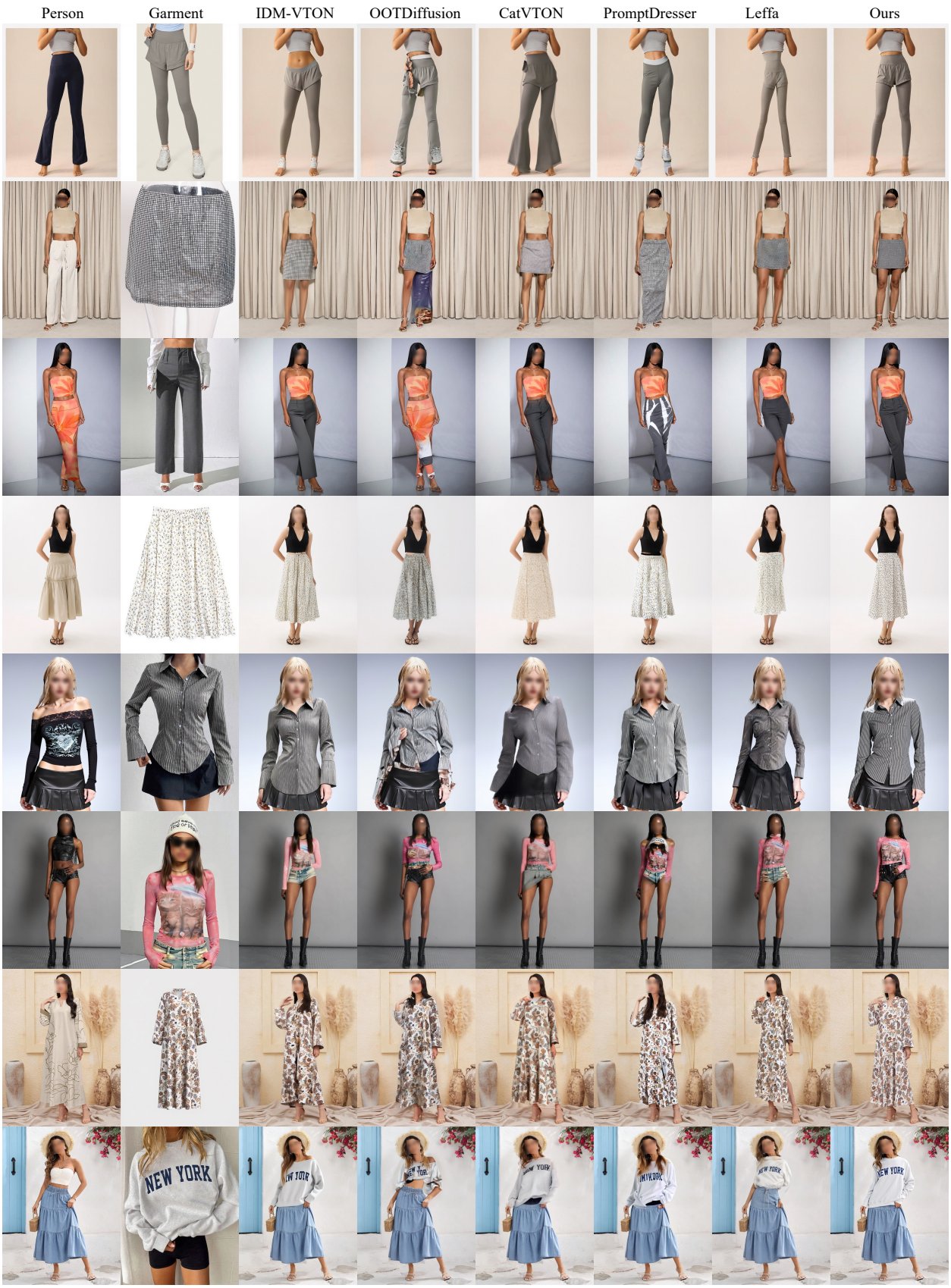

Figure 13. Qualitative comparisons on the in-the-wild images. Please zoom in for more details.

Figure 14. Qualitative comparisons on the in-the-wild images. Please zoom in for more details.

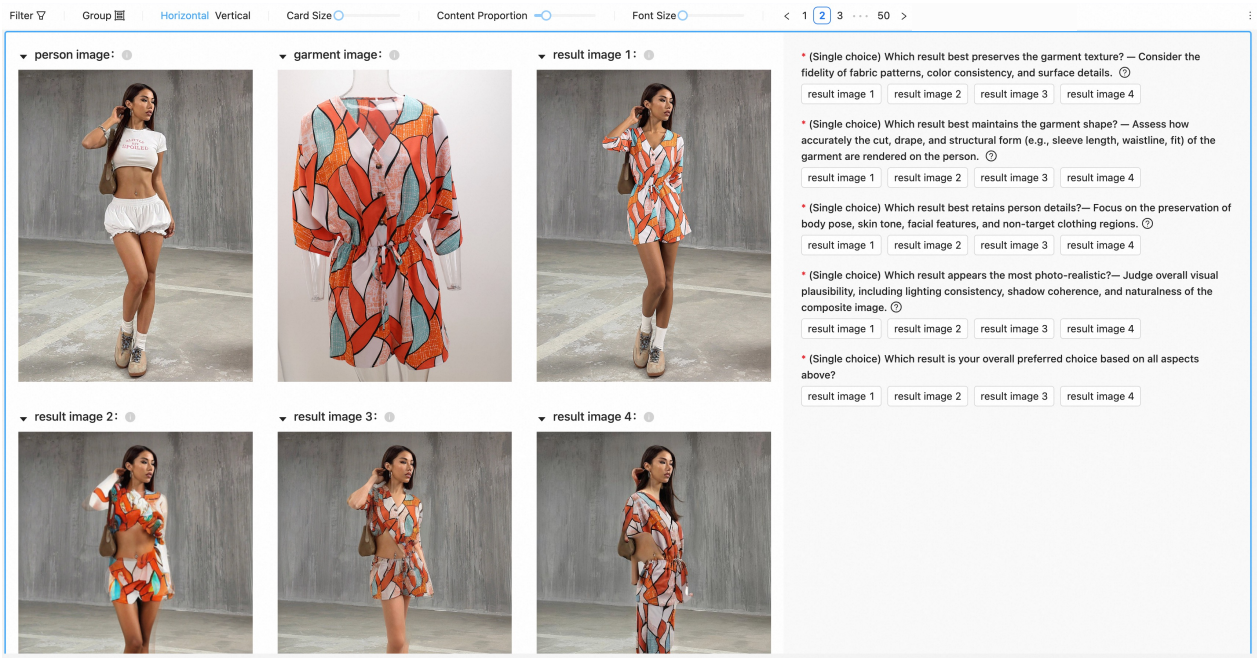

Figure 15. User study interface.