PromptDresser: Improving the Quality and Controllability of Virtual Try-On via Generative Textual Prompt and Prompt-aware Mask

Supplementary Material

A. Details of LMM-driven Virtual Try-on Captioning

We provide detailed explanations of the exemplar datasets, task descriptions, and templates for the categories of upper body, lower body, and dresses in Fig. 11, 12, and 13, respectively. We first gave the LMM model the instruction to identify and list detailed attributes of a given person image, including components, such as the facial expression, skin color, clothing logos. We then selected attributes related to the masked regions of the person image or associated with the style of wearing the clothing, such as pose, hair length, and tucking style. For clothing images, we excluded attributes describing fine details, such as logo shapes or patterns, but instead focused on high-level attributes, such as the clothing category or sleeve.

B. Additional Details on User Study

In our user study, we recruited 40 participants to evaluate the images generated by the baselines across 30 questions. For each question, participants selected the model that best addressed the specified criteria.

For questions 1-25, participants compared the images from multiple datasets. Questions 1–10 featured images from six models (*i.e.*, DCI-VTON, LADI-VTON, Stable-VITON, OOTDiffusion, IDM-VTON, and Ours), based on the VITON-HD and SHHQ-1.0 datasets. Questions 11-25 include three categories of DressCode dataset: upper-body clothing (Questions 11-15), lower-body clothing (Questions 16-20), and dresses (Questions 21-25). For these questions, images from five models were compared, excluding DCI-VTON.

Participants answered the following three questions for each image set:

- Clothing shape: Select the image that best reflects the length and shape of the given garment.
- Clothing details: Select the image that best reflects the text, texture, and pattern of the given garment.
- Overall quality: Select the image of the best overall quality.

For questions 26-30, participants evaluated images generated using the VITON-HD dataset and selected the one that best matched the style described as “untucked, tight fit, and sleeve rolled up.”

Methods	SSIM↑	LPIPS↓	FID↓	KID↓
Ours w/ LLaVA	0.8663	0.1175	8.85	0.91
Ours w/ GPT-4o	**0.8686**	**0.1119**	**8.54**	**0.67**

Table 5. Ablation Results on VITON-HD [10] dataset.

C. Experimental Details

Baselines. We compare our model to four diffusion-based models (LADI-VTON[38], DCI-VTON [18], StableVITON [28], and IDM-VTON [11]). We use pre-trained weights if available; otherwise, we re-implement them using official code. LADI-VTON, DCI-VTON, and StableVITON, all based on Stable Diffusion 1.5, generate images at 512×384 resolution. To ensure a fair comparison, we upscale the outputs to 2× using Real-ESRGAN [50].

Implementation Details. We utilize a frozen SDXL [39] and SDXL inpainting model [26] as the reference and main U-Net, respectively. During inference, we set the denoising step as 30 with σ set to 0.5 for prompt-aware mask generation. To maintain overall pose consistency, we retain hand and foot details within the inpainting mask by Sapiens [27]. Additionally, we use GPT-4o [1] to automatically generate high-quality captions for pre-defined attributes across all experimental datasets.

D. Additional Experimental Results

Comparison to other LMMs. In this paper, we utilize GPT-4o to generate captions for all experimental datasets. To investigate whether our model exhibits a high dependency on GPT-4o in test time, we evaluated it using text prompts generated by an open-source LMM called LLaVA [35]. As shown in Table 5, prompts from LLaVA exhibit slightly degraded performances in the unpaired setting (*i.e.,* FID and KID) but achieve comparable scores in a paired setting (*e.g.,* SSIM and LPIPS), compared to GPT-4o. This demonstrates that the proposed textual prompt can effectively be generated by open-source LMMs such as LLaVA, other than GPT-4o.

Ablation Study on σ. In this paper, we introduce a novel prompt-aware mask to preserve the original person's appearance and enable flexible text-based image manipulation. In generating the mask, we apply early stopping for computational efficiency and adjust the number of inference steps through a hyper-parameter σ. As the value of σ increases, the generation time for the prompt-aware mask decreases. We set the number of denoising steps to 30 across

σ	SSIM ↑	LPIPS ↓	FID ↓	KID ↓
0.8	0.868	0.1122	8.60	0.68
0.7	0.868	0.1119	**8.53**	0.69
0.6	0.868	0.1120	8.55	0.71
0.5	**0.869**	0.1119	8.54	0.67
0.4	**0.869**	**0.1118**	8.54	0.69
0.3	**0.869**	**0.1118**	**8.53**	**0.65**

Table 6. Ablation results for σ values.

all configurations. Table 6 shows the performance behavior based on different σ values. The lowest σ value (0.3) results in more accurate refined masks, achieving the best performance across all metrics. However, slightly reduced performance can be traded off for efficient inference times. In this paper, we adopt $\sigma = 0.5$, which offers inference efficiency while maintaining FID and KID values comparable to those achieved with $\sigma = 0.3$.

Method	sec/image	SSIM ↑	LPIPS ↓	FID ↓	KID ↓
IDM-VTON (40step)	5.84	0.8613	0.1018	9.14	1.18
$\text{Ours}_{\text{pose}}$	5.78	**0.8778**	**0.0967**	9.07	1.16
Ours	5.78	0.8686	0.1119	**8.54**	**0.67**

Table 7. Comparison of IDM-VTON at similar inference time.

Comparison of Inference Time. PMG derives a refined mask from the coarse mask during the initial denoising steps. While this process may introduce additional computational overhead, Table 7 shows that, when compared to IDM-VTON with slightly increased steps, our method demonstrates superior performance in the unpaired setting while maintaining comparable inference time.

	GPT-Human	**Human-Human**
STS	0.8622	0.8889

Table 8. STS between GPT-Human and Human-Human pairs.

Bias in LMM Evaluation. Recent studies [20, 55] validated LMMs for evaluation tasks such as image-to-text and multi-image-to-text alignment. Similarly, we used an LMM to evaluate outfit labeling. To verify the LMM-based evaluation, a user study (Fig. 8(c) in the main paper) assessed text-image alignment via human feedback, showing significant improvement over baselines. To verify LMM reliability, four human annotators labeled 100 VITON-HD test images. Table 8 shows mean semantic textual similarity (STS) between GPT and human labels (GPT-Human), as well as between all human-labeled pairs (Human-Human). Comparable GPT-Human and Human-Human scores confirm GPT's alignment with human judgment.

Additional Visual Results and Diversity. Fig. 9 demonstrates outfit changes and transparent cases generated under various textual conditions, combining complex scenarios such as tops, bottoms, and outerwear. Although trained solely on the DressCode dataset, our method effectively edits images with multi-layered clothing and diverse outfits. We plan to update the image combinations in future versions. Table 9 compares SSIM and LPIPS between tucked and untucked generations. By leveraging rich text prompts and flexible mask, our method achieves higher editability than IDM-VTON.

Figure 9. Multi-layer / transparent outfit generation images.

	SSIM ↓	**LPIPS ↑**
IDM-VTON	0.9401	0.0405
Ours	**0.8702**	**0.1030**

Table 9. Comparison of 'tucked in' vs. 'untucked'.

Figure 10. Generation results of VITON-HD.

Additional Visual Results of Mask Augmentation. Fig. 10 shows the effect of mask augmentation. As noted in the paper, training solely with fine masks causes the model to fit the clothing strictly within the mask region. Without augmentation, it fails to handle coarse masks (*e.g.*, the large rectangular mask), generating overly long garments, as shown in Fig. 10.

Additional Qualitative Comparisons. We present additional qualitative results in Fig. 14 and 15. The first three rows in Fig. 14 depict generated images on the VITON-HD [10] dataset using a model trained on the same dataset, while the fourth and fifth rows show generated images on the SHHQ-1.0 [15] dataset. Our model consistently generates the most realistic images, even for complex poses (rows 1 and 2), and addresses the issue of following the shape of the original clothing (rows 3 and 4). Notably, in the third row, only our model accurately captures the shape of the given cropped top. Moreover, Fig. 15 shows additional results for the upper body, lower body, and dresses categories on the DressCode [37] dataset. Similar to the results on the VITON-HD dataset, our PromptDresser accurately generate the length of the clothing and mitigate the constraint the model follows the original clothing's shape, highlighting the effectiveness of our rich text prompts and a novel mask refinement process.

Additional text-based editing Results. Fig. 16 and 17 demonstrate the text-editing capability of our PromptDresser on VITON-HD and lower body category of the DressCode datasets, respectively. Fig. 16 shows variations in tucking styles on the VITON-HD dataset, where the given clothing is generated based on the text prompts "fully tucked in", "untucked", and "french tucked". Fig. 17 presents variations on the DressCode dataset, including "loose fit," "tight fit," and "pants rolled up." The generated results demonstrate accurate and text-based editing capability of PromptDresser.

System Prompt

You are a fashion magazine director. Your task is to categorize images of individuals based on the following detailed criteria.
Use template to accurately describe and catalog each image.

output :
{
  “body shape” : {“Describe body shape in 1 words.”},
  “fit of upper cloth” : {“Describe fit of upper cloth in 1 words.”},
  “gender” : {“man”, “woman”},
  “hair length” : {“Describe hair length in 3 words.”},
  “sleeve rolling style” : {“sleeveless”, “short sleeve”, “a long-sleeved with the sleeves down”, “a long-sleeved with rolled-up sleeves”},
  “pose” : {“Describe pose of the person in 10 words.”},
  “hand pose” : {“Describe hand pose of the person in 10 words.”},
  “tucking style” : {“Describe tucking style of the person in 2 words.”}
}

Don't output anything else except output template.

Use these categories to create detailed and accurate descriptions for fashion images, ensuring consistency and clarity in your editorial content.

User Prompt

We provided three example images for each criterion.
Few-shot example:

(1) example 1: {
  "body shape" : "slim",
  "fit of upper cloth" : "tight",
  "gender" : "woman",
  "hair length" : "medium-length, to shoulders",
  "sleeve rolling style" : "a long-sleeved with the sleeves down",
  "pose" : "standing straight, neutral expression, looking directly at the camera",
  "hand pose" : "hands relaxed, hanging down naturally at the sides",
  “tucking style” : “untucked”
}

(2) example 2: {
  "body shape": "curvy",
  "fit of upper cloth": "relaxed",
  "gender": "woman",
  "hair length": "medium to long",
  "sleeve rolling style": "a long-sleeved with rolled-up sleeves",
  "pose": "standing slightly angled, looking away from the camera",
  "hand pose": "hands tucked casually into front pants pockets",
  “tucking style” : “fully tucked”
}

(3) example 3: {
  "body shape" : "slender",
  "fit of upper cloth" : "relaxed",
  "gender" : "woman",
  "hair length" : "short",
  "sleeve rolling style" : "short sleeve",
  "pose" : "standing straight, slightly tilted head, confident expression",
  "hand pose" : "left hand by side, right hand relaxed near pocket",
  “tucking style” : “french tucked”
}

Execute one step at a time, making sure to clearly follow the criteria. Categorize the 4th image using every category.

Output

{
  "body shape": "slim",
  "fit of upper cloth": "relaxed",
  "gender": "woman",
  "hair length": "long, shoulder-length, cascading",
  "sleeve rolling style": "short sleeve",
  "pose": "standing slightly angled, playful smile, looking at camera",
  "hand pose": "right hand gently on hip, left hand hanging down",
  "tucking style": "untucked"
},

System Prompt

You are a fashion magazine director. Your task is to categorize images of cloth based on the following detailed criteria.
Use template to accurately describe and catalog each image.

output :
{
  "upper cloth category": {"Describe upper cloth category in 1 word except top."},
  "material": {"Describe material in 1 word."},
  "neckline": {"Describe neckline in 1 word."},
  "sleeve": {"Describe sleeve in 1 word."},
}

Don't output anything else except output template.

Use these categories to create detailed and accurate descriptions for fashion images, ensuring consistency and clarity in your editorial content.

User Prompt

We provided three example images for each criterion.
Few-shot example:

(1) example 1: {
  "upper cloth category": "blouse",
  "material": "mesh",
  "neckline": "crew",
  "sleeve": "long"
}

(2) example 2: {
  "upper cloth category": "camisole",
  "material": "satin",
  "neckline": "v-neck",
  "sleeve": "sleeveless"
}

(3) example 3: {
  "upper cloth category": "blouse",
  "material": "chiffon",
  "neckline": "off-shoulder",
  "sleeve": "short"
}

Execute one step at a time, making sure to clearly follow the criteria. Categorize the 4th image using every category.

Output

{
  "upper cloth category": "t-shirt",
  "material": "cotton",
  "neckline": "crew",
  "sleeve": "short"
}

Template

reference template
"a {upper cloth category}, {material}, with {sleeve}, {neckline}"

main template
"a {body shape} {gender} wears {fit of upper cloth}, {upper cloth category} ({material}), {neckline}, {sleeve rolling style}, {tucking style}. With {hair length} hair, {pose} with hands {hand pose}"

Figure 11. Detailed explanation of the exemplar dataset, task description, and templates for the upper body category.

System Prompt

You are a fashion magazine director. Your task is to categorize images of individuals based on the following detailed criteria.
Use template to accurately describe and catalog each image.

output :
{
 “body shape” : {“Describe body shape in 1 words.”},
 “fit of lower cloth” : {“Describe fit of lower cloth in 1 words.”},
 “length of lower cloth” : {“Describe length of lower cloth in 1 words.”},
 “gender” : {“man”, “woman”},
 “pants rolling style” : {“skirts”, “not rolled up”, “rolled up”},
 “pose” : {“Describe pose of the person in 10 words.”},
 “hand pose” : {“Describe hand pose of the person in 10 words.”},
 “tucking style” : {“Describe tucking style of the person in 2 words.”}
}

Don't output anything else except output template.

Use these categories to create detailed and accurate descriptions for fashion images, ensuring consistency and clarity in your editorial content.

User Prompt

We provided three example images for each criterion.
Few-shot example:

(1) example 1: {
 "body shape" : "slim",
 "fit of lower cloth" : "regular",
 "length of lower cloth": "full",
 "gender" : "man",
 "pants rolling style": "rolled up",
 "pose" : "standing upright, casual stance, facing slightly to left",
 "hand pose" : "left hand holding bag, right hand relaxed at side",
 “tucking style” : “untucked”
}

(2) example 2: {
 "body shape": "slim",
 "fit of lower cloth": "loose",
 "length of lower cloth": "knee",
 "gender": "man",
 "pants rolling style": "not rolled up",
 "pose": "standing upright, facing forward, casual and relaxed stance",
 "hand pose": "right hand relaxed, left hand slightly bent at side",
 “tucking style” : “untucked”
}

(3) example 3: {
 "body shape": "slender",
 "fit of lower cloth": "fitted",
 "length of lower cloth": "mini",
 "gender": "woman",
 "pants rolling style": "skirts",
 "pose": "standing upright, facing forward, slight shift in weight",
 "hand pose": "left hand holding clutch, right hand relaxed by side",
 “tucking style” : “fully tucked”
}

Execute one step at a time, making sure to clearly follow the criteria. Categorize the 4th image using every category.

Output

{
 "body shape": "slender",
 "fit of lower cloth": "loose",
 "length of lower cloth": "short",
 "gender": "woman",
 "pants rolling style": "not rolled up",
 "pose": "standing upright, facing slightly to left, relaxed",
 "hand pose": "hands lightly resting in pockets, elbows bent slightly",
 "tucking style": "French Tucked"
}

System Prompt

You are a fashion magazine director. Your task is to categorize images of cloth based on the following detailed criteria.
Use template to accurately describe and catalog each image.

output :
{
 "lower cloth category": {"Describe lower cloth category in 1 word."},
 "material": {"Describe material in 1 word."},
}

Don't output anything else except output template.

Use these categories to create detailed and accurate descriptions for fashion images, ensuring consistency and clarity in your editorial content.

User Prompt

We provided three example images for each criterion.
Few-shot example:

(1) example 1: {
 "lower cloth category": "jeans",
 "material": "denim",
}

(2) example 2: {
 "lower cloth category": "shorts",
 "material": "cotton",
}

(3) example 3: {
 "lower cloth category": "skirts",
 "material": "polyester",
}

Execute one step at a time, making sure to clearly follow the criteria. Categorize the 4th image using every category.

Output

{
 "lower cloth category": "shorts",
 "material": "cotton"
}

Template

reference template
 "a {lower cloth category}, {material}"

main template
"a {body shape} {gender} wears {fit of lower cloth} with {length of lower cloth}, {lower cloth category} ({material}), {pants rolling style}, {tucking style}. {pose} with hands {hand pose}"

Figure 12. Detailed explanation of the exemplar dataset, task description, and templates for the lower body category.

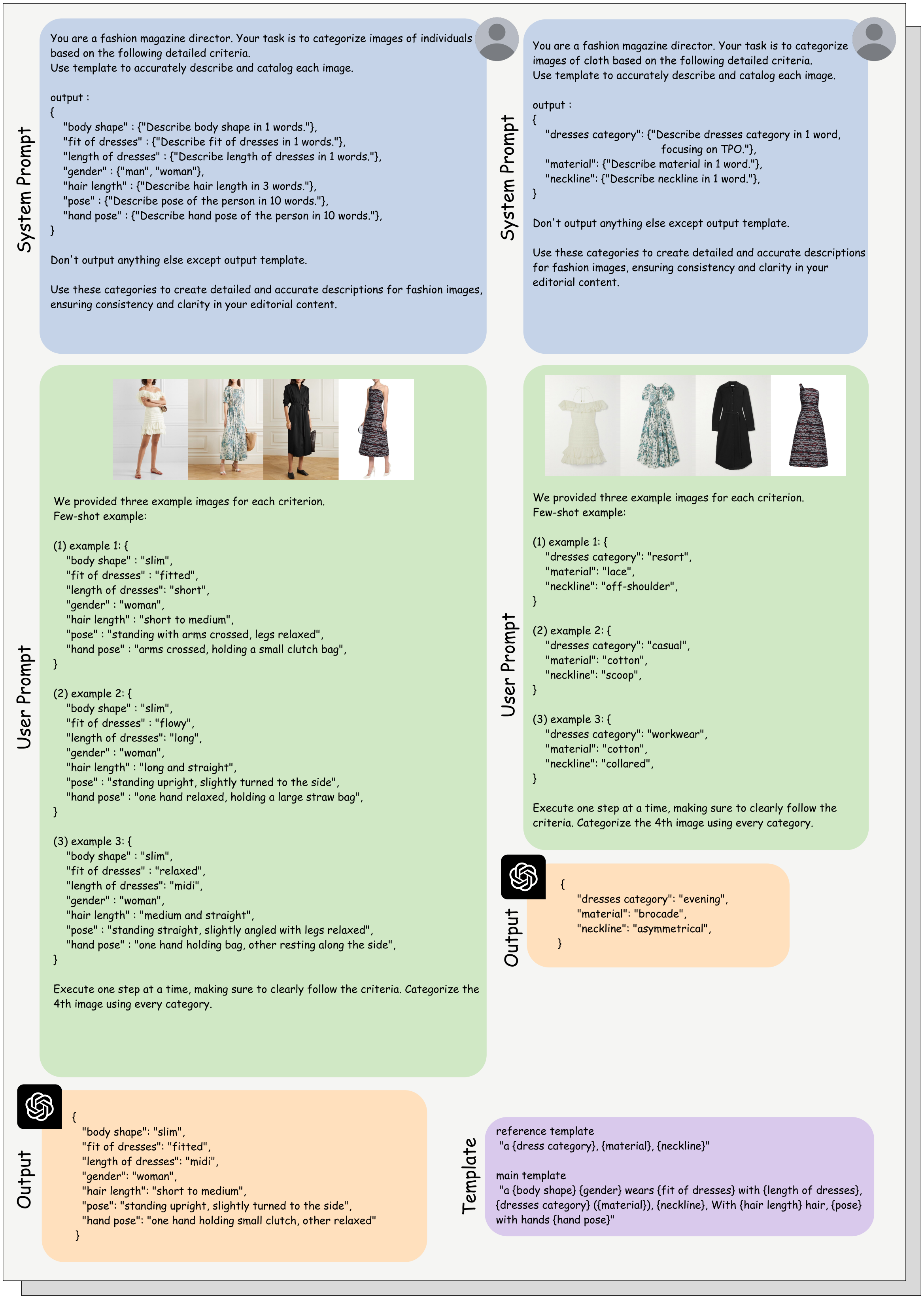
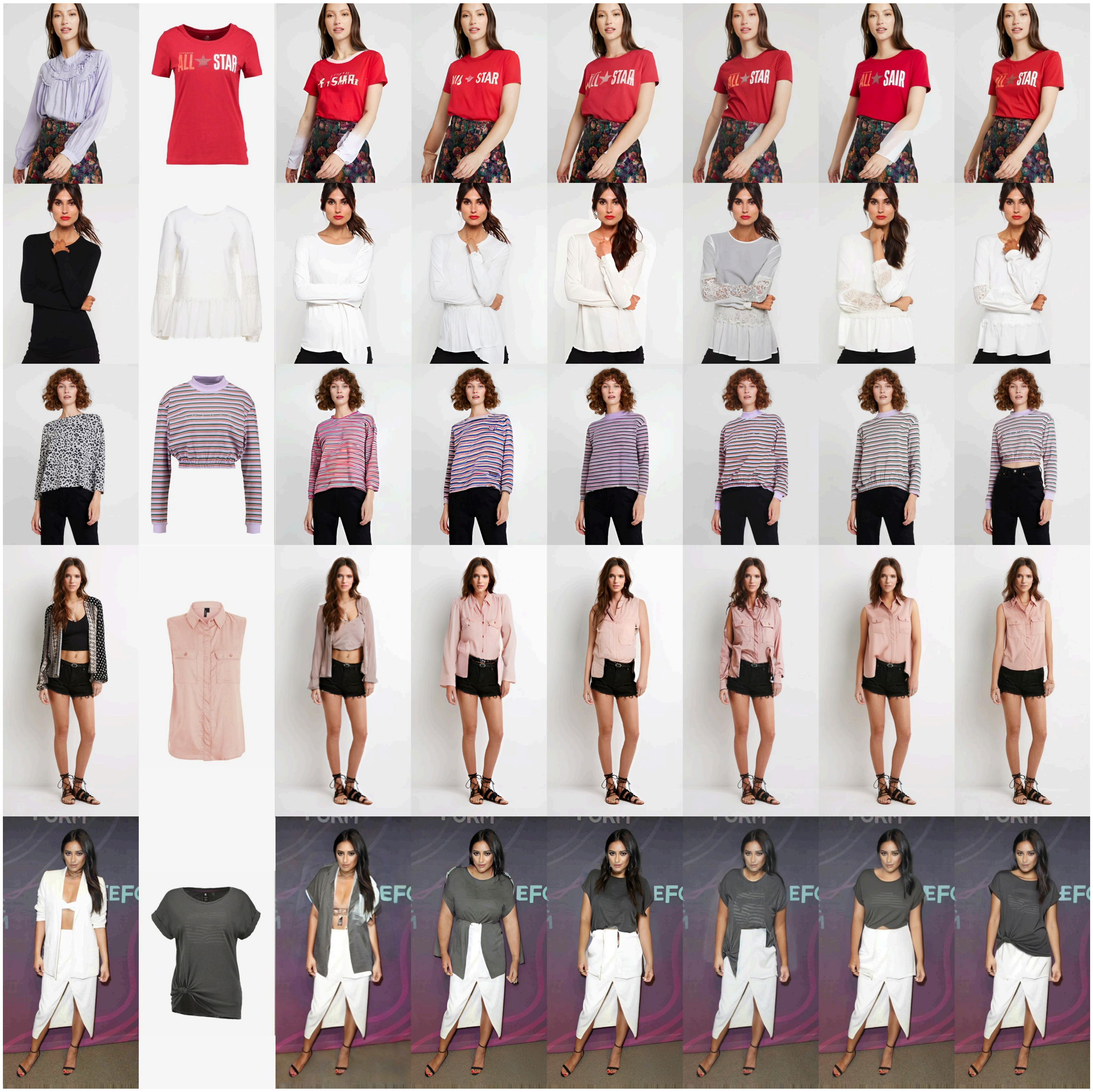


Figure 13. Detailed explanation of the exemplar dataset, task description, and templates for the dresses category.



Target Person | Clothing | LADI-VTON | DCI-VTON | StableVITON | OOTDiffusion | IDM-VTON | Ours

Figure 14. Qualitative comparison with baselines trained on VITON-HD dataset (first / second / third row: VITON-HD, fourth / fifth row: SHHQ-1.0)

Figure 15. Qualitative comparison with baselines trained on DressCode dataset.

Figure 16. Additional text-based editing results for the upper body category of the VITON-HD dataset.

Figure 17. Additional text-based editing results for the lower body category of the DressCode dataset.